Aalto University
School of Science

Juuso Ilomäki

Recognizing Transportation Modes by Multivariate Clustering of Accelerometer Data

In Espoo 14.6.2016
Supervisor: Professor Ahti Salo
Instructor: D.Sc. (Tech.) Tapio Soikkeli

AALTO-YLIOPISTO TEKNIIKAN KORKEAKOULUT PL 12100, 00076 Aalto http://www.aalto.fi	DIPLOMITYÖN TIIVISTELMÄ

Tekijä: Juuso Ilomäki

Työn nimi: Liikkumismuotojen tunnistaminen kiihtyvyysanturidatan ryhmittelyanalyysillä

Korkeakoulu: Perustieteiden korkeakoulu

Koulutusohjelma: Teknillinen fysiikka ja matematiikka

Pääaine: Systeemi- ja operaatiotutkimus | Koodi: F3008

Työn valvoja: Professori Ahti Salo

Työn ohjaaja(t): TkT Tapio Soikkeli

Älypuhelin on laite, joka sisältää lukuisia datan keräämiseen soveltuvia antureita ja joka kulkee lähes aina mukana. Älypuhelimen hyödyntämisestä datan keräämisen työkaluna onkin tullut kiehtova ajatus erilaisten ”lifelogging” ja ”biohacking” -harrastusten kasvattaessa suosiotaan. Ihmiset ovat alati kiinnostuneempia päivittäisistä aktiviteeteistaan numeerisen datan valossa ja mukana kulkeva laite on näppärä tapa tuottaa sitä.

Tässä työssä kerättiin kiihtyvyysanturidataa yhdeksältä testihenkilöltä yhdeksällä laitteella pyrkimyksenä löytää tilastollisiin monimuuttujamenetelmiin pohjautuva menetelmä datan ryhmittelemiseksi liikkumismuodon mukaisiin ryhmiin. Jokainen testihenkilö kantoi laitetta mukanaan noin kolmen päivän ajan ja kirjasi samalla ylös suorittamansa aktiviteetit käyttäen Aalto-yliopistossa kontekstidatan keräämiseen kehitettyä Contextlogger3-ohjelmaa. Ryhmittelyn onnistumista arvioitiin vertaamalla algoritmin tuottamia ryhmiä siihen, mitä käyttäjät olivat kirjanneet.

Työn tuloksena selvisi, että valitulla lähestymistavalla kävelydatat saatiin erottumaan varsin hyvin muista liikkumismuodoista, mutta erilaisista moottoriajoneuvoista lähtöisin olevia datoja ei kyetty kunnolla erottelemaan toisistaan.

Päivämäärä: 14.6.2016	Kieli: Englanti	Sivumäärä: 52

Avainsanat: Kontekstin tunnistaminen, liikkumismuodot, anturidata, ryhmittelyanalyysi

AALTO UNIVERSITY
SCHOOLS OF TECHNOLOGY
PO Box 12100, FI-00076 AALTO
http://www.aalto.fi

ABSTRACT OF THE MASTER'S THESIS

Author: Juuso Ilomäki

Title: Recognizing Transportation Modes by Multivariate Clustering of Accelerometer Data

School: School of Science

Degree Program: Engineering Physics and Mathematics

Major: Systems and Operations Research | Code: F3008

Supervisor: Professor Ahti Salo

Instructor(s): D.Sc. (Tech.) Tapio Soikkeli

The dawn of the Quantified Self movement suggests that people are increasingly interested in collecting and analyzing data from their everyday lives. Data can be recorded by using specially designed 'life logging' devices, but a regular smart phone suits for the task as well. Smart phones are often carried in a pocket throughout the day and have various built-in sensors which can be harnessed for data collecting.

In this thesis we collected smart phone accelerometer data from nine test subjects, aiming to be able to algorithmically cluster the data in transportation mode -specific groups. The test subjects logged their daily activities by using Contextlogger3-application, which has been developed at Aalto University for this specific purpose. Performance of the algorithmic clustering was later evaluated by comparing the formed groups to what the test subjects had logged in reality.

We were able to recognize walk data fairly accurately, but with the selected approach it was impossible to separate car and bus data.

Date: 14.6.2016 | Language: English | Number of pages: 52

Keywords: Context recognition, mode of transportation, sensor data, clustering

Acknowledgements

I got the spark for the thesis in course *Tilastolliset Monimuuttujamenetelmät* where both I and Kimmo Karhu from SoberIT attended in the spring of 2012. Kimmo informed me about their current work on smart phone context logger application which was to be used for collecting data about various daily activities. Collected data would later be used in a variety of studies.

One of the planned studies was recognizing the mode of transportation, for which they were actually seeking a thesis worker who - if possible - could be an applied math major. Luckily for me, I was at the same time an applied math major seeking a topic for my thesis. We agreed that Kimmo would later send me some more info about their project and the rest is history.

Working with Kimmo as my superior and Tapio Soikkeli as a thesis instructor was very enjoyable time. At first it was just experimenting with the nascent context logger software, but soon thesis structure and scope started to get clear and I asked Professor Ahti Salo for supervisor. Thesis topic was confirmed soon afterwards.

I want to thank Kimmo, Tapio and Ahti for crucial help with the thesis. Thanks also for Professor Esa Saarinen for inspiration and helping to increase my study motivation which had gotten a bit low at one point.

Espoo 14.6.2016

Juuso Ilomäki

Juuso Ilomäki

Table of Contents

Notations and Abbreviations

FFT	Fast Fourier Transform
GPS	Global Positioning System
API	Application Programming Interface
IoT	Internet of Things
Contextlogger3	Context logger application built on top of Funf framework. Developed in Aalto University for recording context specific data from Android devices.
Funf framework	Open Sensing Framework for mobile devices maintained by Behavio (http://www.behav.io/). The core concept is to provide an open source, reusable set of functionalities enabling the collection, uploading and configuration of a wide range of data signals accessible via mobile phones.
NCSS	Statistical software for analyzing data
Instance of data	Activity-specific dataset collected by one test subject by tagging start and end markers for one activity. Instances are divided into shorter batches of data.
Batch of data	A ten-second dataset consisting of 500 consecutive accelerometer readings

1 Introduction

Smartphones and similar mobile devices have become a ubiquitous part of modern everyday life. First the cell phone which became an omnipresent gadget inside people's pockets but evolution was fast. Ever increasing number of features induced a transformation process: The device we had at first considered "just a phone" was suddenly becoming much more. It was not just for phone calls anymore. A legendary game 'Snake' along with many others came out and SMS's started flying around. There were early versions of mobile calendars and notepads, but usability was limited because of small screen and numpad interface. When the devices with large touch screens started to roll out it was time for a revolution – the smartphone revolution.

Big touch screen, computer-like operating system, various sensors and endless number of both manufacturer and user developed applications are able to deliver almost desktop computer level user experience. Many applications, such as calendar and e-mail, can in fact be synchronized with their desktop counterparts. User interface is tilted so that "down is always down", no matter which way the user holds the device. With few swipes of a finger people can share their pictures, activities and locations in social media. Designated applications make it ridiculously easy to keep track on performed sports activities while plotting jogging routes on a map can happen automatically.

What lies behind of smartphone's impressive capabilities is, besides the ever increasing amount of raw computing power, a number of built-in sensors. Smartphone is a self-sensing gadget aware of its surroundings. For example, tilting the user interface based on how the user holds the phone is possible because of acceleration sensors sensing the earth's pull. The phone is aware of its orientation towards earth.

Accelerometer is the single most important sensor from this thesis' point of view, but there are some other potentially useful sensors for activity recognition. Magnetic field sensor is one, but GPS locator could be even more interesting. Besides providing the location information, GPS data would also provide a straightforward way to use velocity in the analysis. Knowing the velocity would, at least intuitively, be very helpful with some modes of transportation. Even though we accessed accelerometer data, it was not possible to calculate velocity from it because we would, for example, need to know the initial velocity in the beginning of the data – we will come back to this in the discussion.

Despite the potentially better activity recognition when combining accelerometer data with GPS, we wanted to only use accelerometer for couple of reasons. First, it is possible to maintain better privacy level when the location information is not revealed. We wanted to figure out how reliable recognition can be achieved without revealing the location. Second, GPS would need a direct access to satellites via open sky whereas accelerometer is useful also in tunnels and indoors. We wanted a robust method which could possibly be used in subway and indoors.

We collected data from nine test subjects who each used their own device. For a time span of few days the test subjects let the device know when they started and stopped certain activities, such as walking, so it was later possible to catch these *instances* of data by querying the user-added "timeline tags". For collecting the data we used Contextlogger3 (Chaudhary, 2013; Mannonen et al., 2013), a software developed in Aalto University basing on Android-compatible Funf-framework (Aharony et al., 2011).

Data contained three dimensional (x, y, z) accelerometer readings at a rate of 50 readings per second. We cut the recorded instances to 500-reading *batches* and calculated certain descriptive statistical numbers to represent the batches in clustering. For analyzing the data we used NCSS. We explored different algorithms and different parameters with an objective to find such multivariate method which could be used to identify different modes of transportation.

1.1 Motivation

Quantified self is a buzzword or – some say – even a movement (Swan, 2013) to incorporate technology for acquiring data of a person's daily life. People use various data collecting devices to measure and self-monitor their activities and use this data in life logging, body hacking and self-quantifying. There are special devices and wearable sensors for data collecting, but data can also be collected by using general devices such as smartphones. Using smartphones is attractive because then there would be no need to carry extra devices.

If it is possible to create a model which can accurately give interpretations about current mode of transportation, it would become fairly easy to collect big data of transportation habits of individuals or groups of people. Most smartphones can collect at least: Accelerometer data, GPS-data, magnetic field data, cell tower ID, camera data, microphone data and software logs. Accelerometer and GPS are likely two of the most potential sensors from the transportation mode recognition point of view. If the task can be accomplished without GPS, the method would be more robust and better privacy level could be maintained. Recognizing the mode of transportation by using cell phone data is a research problem that both universities research and firms are trying to solve. For example Google has rolled out an activity recognition API for Android devices in 2015.

1.2 Research Questions

The purpose of this thesis was to explore whether it is possible to use multivariate analysis for identifying the mode of transportation when only cell phone acceleration data is available. Many similar studies have been conducted previously, but methods for recognition as well as activities to be recognized vary from one study to the next. Many studies also have a strictly defined setup where test subjects wear specific sensors on specific body parts.

Using GPS data could make the task easier because it makes it possible to know the velocity of the phone and location data can be used to narrow down the possible transportation modes. We sought to recognize these modes based on accelerometer only which

would be more robust method because it would also work without connection to satellites and would also ensure the privacy for users.

The main research objective was to build a framework based on multivariate clustering which would separate data from transportation mode specific origins into their own groups. It was also studied what kind of matters should be taken into consideration when collecting the data for this type of model.

1.3 Structure of the Thesis

The structure of the thesis is as follows. Chapter 2 covers the background research for the thesis. Chapter 3 describes the data collecting process and the methods used when pre-processing the data for multivariate analysis. Chapter 4 presents the research results. The results section is divided into two halves. First, various clustering methods are applied to a subset of collected data and then, in the second part, the best method is applied to the full data set. Chapter 5 presents the conclusions and discusses difficulties faced when building the transport recognition framework. Furthermore, propositions for similar future studies are presented.

2 Literature Review

In recent years, people have become increasingly interested in collecting and analyzing numerical data about their daily activities. The continually growing computational capacity of devices and growing capacity of network bandwidth along with explosively growing number of sensors connected to Internet have created a new emerging ecosystem: Internet of Things or IoT. This IoT has been an important enabler to what is referred even as Quantified Self movement where people incorporate technology to collect and analyze various data about their daily activities. (Atzori et al., 2010; Swan, 2013, 2012)

Accelerometer data has previously been used in various studies to activity recognition. Some studies have used specifically designed sensors attached on various body parts (Bao and Intille, 2004) and focus might then be more in gesture recognition. Bao and Intille used sensors located on the subjects' hip, wrist, arm, ankle and thigh and achieved an overall 84.26% recognition rate on activities such as walking, running, reading, climbing stairs.

Then there are also studies where only one single sensor is used for data collecting. This single sensor could be specifically designed accelerometer (Randell and Muller, 2000; Ravi et al., 2005) or it can be general device such as smartphone (Lee and Cho, 2011; Zhang et al., 2010). Bao and Intille achieved fairly good recognition with multiple sensors, but good results have also been achieved by using a single sensor (Long et al., 2009). Long et al. used a single accelerometer attached to test subjects' waists and achieved classification accuracy of about 80% when recognizing activities such walking, running, cycling, driving and sports.

Specific wearable sensors are not necessarily needed for data collection. The fact that smartphones have become a device many carry with them throughout the day (Iftode et

al., 2004) and the fact that the smartphones have a number of sensors along with other means (e.g. software logs, microphone) for data collecting makes it fairly easy to collect plenty of data from either a single person's or from a group's daily activities by using the mobile device as a data collecting tool (Phithakkitnukoon et al., 2010). Obstacle is not the lack of infrastructure but rather the quality of the possibly noisy data and also privacy issues (Lane et al., 2010). Smartphones can even be used as a gesture based input device by waving them and interpreting the patterns from data created by built-in accelerometer (Ballagas et al., 2006).

An interesting study from the viewpoint of this thesis was conducted by Kwapisz et al. (2011). The activities they tried to recognize were: walking, jogging, going upstairs, going downstairs, sitting and standing – not any motorized vehicles, but still somewhat similar activities with our case. Kwapisz et al. used Android smartphones as data collecting tool instead of specific accelerometer sensors. Their test subjects carried the devices in their pockets and did not get any specific training on how to perform the activities. They simply carried the device in their pocket and acted naturally. Devices recorded 10 second batches which each had 200 readings in them.

There are a number of studies where accelerometer data is used to recognize various activities. Recognizing the mode of transportation is a common research problem (Reddy et al., 2010, 2008), but some studies also try to find ways to recognize various movement related bodily activities such as jogging, walking, climbing stairs etc. (Kwapisz et al., 2011) or various daily activities such as dish washing, making bed etc. (Bouten et al., 1997).

Methods for recognition also vary from study to study. Multivariate analysis is used in some studies (Yi et al., 2005), but there are studies which utilize e.g. the Hidden Markov Model (He et al., 2007) or neuro-fuzzy classifiers (Yang et al., 2007). Data might also be preprocessed depending on the method used. For example some studies have used Fourier transformations for helping to catch repeating patterns from the data (Ward et al., 2006). A fine study about various data preprocessing techniques has been conducted by Figo et al. (2010). They suggest that for example mean, median, standard deviation, min and max are suit well for representing data in statistical analysis in this type of study.

Huynh and Schiele (2005) analyzed pre-recorded data by applying clustering analysis and evaluated the performance of individual features for activity recognition. Thus the study objectives were also quite similar to our case. However their data was somewhat different. Huynh and Schiele used the prerecorded data of roughly 200 minutes which had been recorded by two subjects unfamiliar with the researchers. The subjects had been given a script containing various activities which they performed. Data was collected with integrated sensor board attached to the backpacks they were carrying. The setup which included specifically designed sensor board and pre-defined script for subjects on how to perform the activities, created fairly strictly defined data collecting environment. We collected data from the test subjects' everyday life, so the environment in our case was more loosely defined, but in terms of methodologies the study was quite similar to our case. Huynh and Schiele also applied FFT for their data in preprocessing, which seems a good approach considering their fairly strictly prepared data collecting setup.

When cell phone data is collected by humans, especially in an uncontrolled environment, the amount of noise varies depending on how the recorded activities are performed. People have individual habits to use their phones; some hold it in their pocket while others may place it on the dash board in a car, for instance. Noise can be reduced by training the test subjects to collect the data in some specific way. The more training and rules are given, the closer the data can get to what is defined to be 'pure activity'. While for example Huynh and Schiele had ready-made scripts for their test subjects there are also a number of studies with much more relaxed conditions (Bao and Intille, 2004).

As mentioned in introduction, a smartphone is a device with a number of built-in sensors suitable for collecting various data and accelerometer is only one commonly used option. The GPS locator is often used for enhancing the recognition (Kantola et al., 2010). While it can enhance the accuracy of recognition, a GPS also has some downsides. First, it needs a direct access to satellites and second, it reveals the accurate location of the user. Having privacy from this kind of tracking is seen even as a basic human right by many (Beresford and Stajano, 2003). When using only accelerometer data, the exact location of user is not revealed and the analysis is based purely on movement which obviously increases privacy.

Organizing data into groups and creating taxonomies is a fundamental mode of learning and understanding. Clustering is one commonly used method for this type of data classification and it has been used in a variety of fields. Clustering is not one specific algorithm but rather a task to be accomplished and it can be achieved by a number of specific algorithms such as K-means (Jain, 2010). Other commonly used clustering algorithms include the nearest neighbor method, the furthest neighbor method, the centroid method, the group average method and the Ward's minimum average method. K-means is what is called a non-hierarchical method, whereas the rest of the methods are hierarchical methods. The main difference between these types is that non-hierarchical methods do not make any a priori assumptions about the number of clusters.

3 Data and Methods

Based on the literature review and the available resources and competencies the following approach was chosen for the study. We would collect accelerometer data from a number of test subjects by equipping them with Android based smart devices with built-in accelerometers which would then be used to record the test subjects' daily activities for a time span of few days.

The test subjects were instructed to carry the devices just like they would do in everyday life. Only instruction given before the recording was how to use the context logger software.

We then applied various multivariate clustering methods on the collected data in order to form transportation mode based clusters. One of the study's purposes was also trying to recognize possible problems in analyzing data originating from everyday context in contrary to more of a laboratory environment. Thus some human decision making was applied to evaluate the reliability of data before the actual clustering analysis.

3.1 Overview of the Data Collecting System

Basically all smartphones have a built-in accelerometer. It was therefore possible to equip a number of test subjects with smartphones which they carried for a time span of few days and logged various daily activities. To record the data we used context logger application – a software designed in Aalto University and designed for this purpose.

By using the context logger, the test subjects were able to place tags in timeline so it would later be possible to catch the accelerometer data of specific activities from the data. For example, a test subject could have tagged that they were travelling in the bus between

time stamps A and B and that they were walking between the time stamps C and D. The same data collecting framework, and the same data was used also in some other studies besides this thesis. Because of this the data also contains some tags not related to transportation activities. An example of the collected data from one person is in Figure 1.

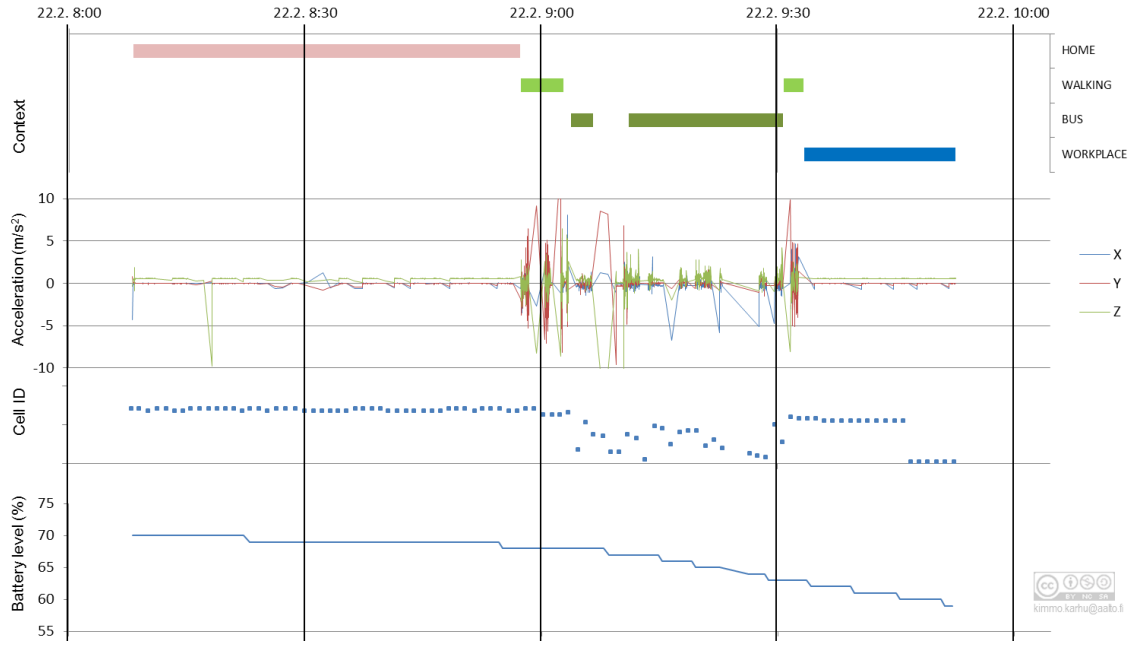

Figure 1: An example of data collected

3.2 Preparing the Data for Multivariate Analysis

The data was first collected into device specific databases. These databases then needed to be extracted from the devices to a form where the actual multivariate analysis could be performed in NCSS.

The collected data was in an encrypted form and the first step in extracting the data was to decode the encryption and to parse separate database files into single csv-files. For this we used python scripts provided by Funf framework (Aharony et al., 2011). The data consisted of accelerometer readings in three orthogonal dimensions (x, y and z) which were collected around 50 times per second.

The next step in preparing the data was to arrange the readings to correct order in timeline, which they were not in the original database files. It was also necessary to scrap all the

excess data not related to transportational activities. This arranging and scrapping was done by using SQL queries in MS Access.

Because of statistical nature of the study we decided to use batches of exact 500 consecutive readings (around 10 seconds measured in time) for the multivariate analysis. Batches were collected randomly by using the following guidelines:

- The same readings were never used in two or more batches.
- Because of the limitations set by the devices there were some gaps in the data which were rejected from the batches.
- The very beginning and very end of collected data sets were rejected because of high accelerations caused by using the device when placing start and stop tags for the activities.
- When the recorded activity had some obvious impairment, the data was not used. Reasoning for all this type of scrapping is discussed in chapter 3.4 Characterizing the Collected Data.

Data contained acceleration readings in three dimensions: X, Y and Z. For the analysis the vector sum of these was calculated as

$$SUM = \sqrt{X^2 + Y^2 + Z^2}. \quad (1)$$

3.3 Clustering Method

Clustering is a multivariate method for grouping a set of data into clusters which have more similar properties within each other than to those of other clusters. Clustering analysis is not a specific algorithm, but, rather, a general task to be solved and it can be achieved by various algorithms. The task was to form transportation specific clusters so that walking would end up having its own cluster, car its own cluster and bus its own cluster. Some data from also few other modes of transportation was available and it was analyzed, but not as systemically as walk, bus and car data.

Clustering algorithms can further be divided into hierarchical and non-hierarchical categories. In hierarchical clustering each measurement point forms its own cluster in the

beginning and these clusters are combined step-by-step until a satisfactory result is achieved. Hierarchical methods rely on human decision making and work well when there is not information available about the final number of clusters.

3.3.1 Hierarchical Methods

In this thesis we compared five different hierarchical methods for recognition: Nearest Neighbor Method (4.1.1), Furthest Neighbor Method (4.1.2), Centroid Method (4.1.3), Group Average Method (4.1.4) and Ward's Minimum Average Method (4.1.5). The task was to group *n* observations into *g* groups.

A pivotal part of hierarchical methods is called dendrogram, which is a tree-like graph where stem is in one side of the graph (in this thesis right) and branches in the other side (in this thesis left). When moving from left to right the observations are annexed one by one into larger groups. The length of a single branch describes how much it differs from the other branches.

In all hierarchical methods the basic principle is the same, and only certain variables differ between the methods (see **Error! Reference source not found.** for the list of variables).

Method	**Variables**
Nearest Neighbor Method	$\alpha_k = \alpha_l = 1/2, \beta = 0, \gamma = 1/2$
Furthest Neighbor Method	$\alpha_k = \alpha_l = 1/2, \beta = 0, \gamma = -1/2$
Centroid Method	$\alpha_k = n_k/n_r$, $\alpha_l = n_l/n_r$, $\beta = - \alpha_k \alpha_l$, $\gamma = 0$
Group Average Method	$\alpha_k = n_k/n_r$, $\alpha_l = n_l/n_r$, $\beta = 0, \gamma = 0$
Ward's Minimum Average Method	$\alpha_k = (n_k+n_s)/(n_r+n_s)$, $\alpha_l = (n_l+n_s)/(n_r+n_s)$, $\beta = -n_s/(n_r+n_s)$, $\gamma = 0$

Table 1: Variables used for calculating new distance matrix depend on the method used

The basic algorithm can be presented as follows:

1) Start from the initial clustering C_n in which each observation forms its own cluster and form the corresponding distance matrix.
2) Form new clustering C_j based on the previous clustering C_{j-1} by annexing the two currently closest groups while other groups are unaffected.
3) Form the corresponding distance matrix.

4) Repeat steps 2 and 3 until a satisfactory result is achieved.

The distance matrix is formed as follows. Let groups *k* and *l* be the closest with the distance d_{kl} between them. We then assume that group *k* has n_k members and group *l* has n_l members. New distance matrix can be formed as

$$d_{rs}^2 = \alpha_k d_{ks}^2 + \alpha_l d_{ls}^2 + \beta d_{kl}^2 + \gamma |d_{ks}^2 - d_{ls}^2| \quad (2)$$

The same equation can be applied to all hierarchical methods used and only the variables α_k, α_l, β and γ (see **Error! Reference source not found.**) depend on the method used.

When evaluating mathematical performance, the cophenetic correlation coefficient can be used as a metric. Cophenetic correlation coefficient is Pearson product-moment correlation between the real distances of observations and distances of grouping method specific distances. The bigger the cophenetic correlation values the better.

Another useful value to evaluate mathematical performance of clustering is delta coefficient (see equation 3).

$$\Delta_a = \left[\frac{\sum_{k<l}^{n} |d_{kl} - {d_{kl}}^*|^{\frac{1}{a}}}{\sum_{k<l}^{n} ({d_{kl}}^*)^{\frac{1}{a}}} \right]^a \quad (3)$$

Where a = ½ or a = 1 depending on which delta value is calculated. d_{kl} is the original distance between the two points and d_{kl}^* is Euclidian distance between same points after the clustering. The smaller the delta values the better the clustering.

We used NCSS software to perform iterations. NCSS also provided the dendrogram graphs which are presented in the results section. Dendrogram is a tree diagram where the "leaves" represent the observed data which is then merged into branches based on the similarity of the observations.

Evaluating the results provided by different methods needs also some human reasoning, because the methods do not have any knowledge about the real origin of various observations. Cophenetic correlation and delta values can be used for evaluating only the mathematical performance.

3.3.2 K-means

The non-hierarchical method used for clustering in this thesis is called k-means first developed by J.A. Hartigan and M.A. Wong of Yale University (Hartigan and Wong, 1979). Non-hierarchical clustering consists of the following parts:

1) Choose the number of clusters g.
2) Form the initial clusters by attaching each observation to some cluster.
3) Increase the homogeneity of clusters by moving observations between the clusters.
4) Iterate phase 3 until it is no longer possible to increasing the homogeneity.

Again as with the hierarchical methods, NCSS could be used for all time consuming iteration work.

NCSS lets the user choose the following options in k-means:

- Minimum and maximum number of clusters sets the range of clusters to try, although k-means algorithm finds a cluster configuration for a fixed number of clusters.
- Reported clusters specify the number of clusters in the final solution.
- Random starts sets the amount of initial configurations to try. The k-means algorithm finds a local optimum so larger amount of random starts helps to find global optimum.
- Max iterations specifies the maximum number of retries before the algorithm is aborted.
- Percent missing option specifies the percentage of missing values allowed before an observation is skipped.

3.4 Characterizing the Collected Data

There were nine test subjects who each had their own device for data collecting. The test subjects carried the devices for 1-3 days, varying from person to person, and recorded their daily activities by using specifically designed context logger software (Chaudhary, 2013; Mannonen et al., 2013). Same data was also used for other studies and contained also non-transportation based activity tags. All subjects still had at least one instance of transportation based data recorded.

In this thesis *walk*, *bus* and *car* were in spotlight because most of the transportation based recordings originated from those activities but all transportation related data was still picked for analysis.

It turned out that some of the data was corrupted and had to be abandoned. Still, more than enough was available for the purpose of this thesis. Corrupted data is discussed and corrections/improvements for further context logger and/or mode of transportation research are suggested. It was fairly easy to identify the corrupted data and to avoid it in analysis. This, however, caused some extra manual work in the form of evaluating the validity of data before performing the actual analysis.

Few days of data from each subject was separated into *instances* of walking, driving a car, travelling by bus and also to few other transportation related activities. Because the activities were from subjects' everyday life, they were of varying lengths. Because the approach based on statistical analysis, the recordings were chopped into exact 500-reading *batches* to keep them commensurate. A context logger was set to collect at maximum possible frequency, around 50 readings/second, and thus each batch was around 10 seconds measured in time. Around 10 second time frame was seen suitable in the light of similar previous research, for example Kwapisz et al. (2011).

Data was named in the following format: <subject> - <activity> - <instance> - <batch> so for example a notation G-Bus-1-02 indicates that the dataset in question is second batch from subject G first bus instance and B-Walk-2-05 is fifth batch from subject B second walk instance. If the whole instance is referred to, then the batch number is simply left out from the end, for example, G-Bus-1. The following chapters describe the collected data subject by subject.

Coincidentally most of the corrupted data seemed to originate from the devices A-E whereas the data collected from the devices' F-I seemed to be mainly legitimate. When the data is introduced in alphabetical order it might at first look like barely any legitimate data was available. While the amount of corrupted data was unfortunate, we still had plenty of legitimate data available.

3.4.1 Subject A

Subject A had 1 instance of bus data, 4 instances of walk data and 2 instances of cycling data. Unfortunately all walk data and all cycling data was corrupted. An example of corrupted cycling data can be seen in the figure below (Figure 2). Bus data from A still seemed to be fine so it was chopped into 4 batches which was maximum amount possible for this rather short instance.

Bus data was the last one recorded by A so it is possible that old or incompatible version of Android and context logger versions were used for the first activities and the version was updated before recording the bus instance. Please note that, unlike in the rest of the figures, in Figure 2 all acceleration components (X, Y and Z) are presented because this better illustrates the problem with the data. Single components had acceleration profiles too cyclic to be credible. We had seen similar phenomenon happening with test data when the old version of context logger was used. Later in the thesis only acceleration SUM vector is presented unless there is some specific reason to demonstrate separate components.

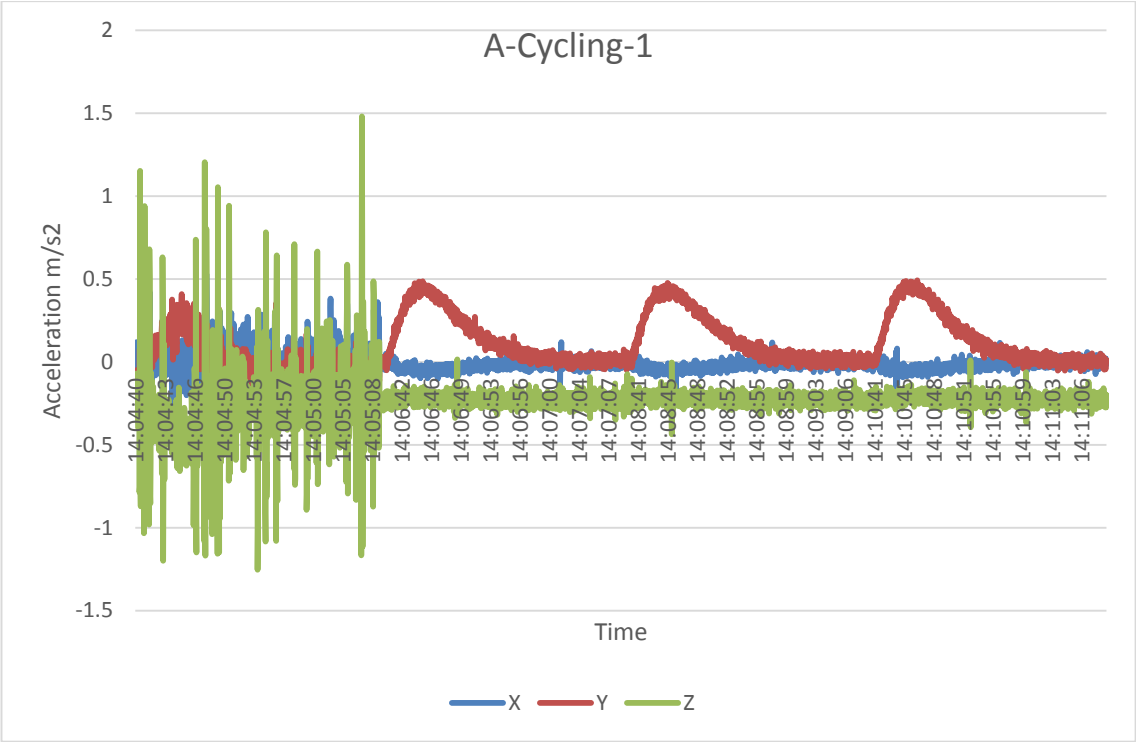

Figure 2: Subject A's cycling data showed constant acceleration to certain directions and also too regular rhythm to be trusted. Similar problem was also found in A's walk data.

3.4.2 Subject B

Subject B had 1 instance of car data and three instances of walk data. B-walk-2 and B-walk-3 likely had actual walk data only at the beginning of the instances and superfluous ambient data at the end. It is possible that the subject forgot to put closing tag at the end of these walk instances.

Everything except ambience at the end of instances was chopped into batches and used in analysis. Ambient data from the end of these instances was obviously left out. An example of B-walk data can be seen in the Figure 3.

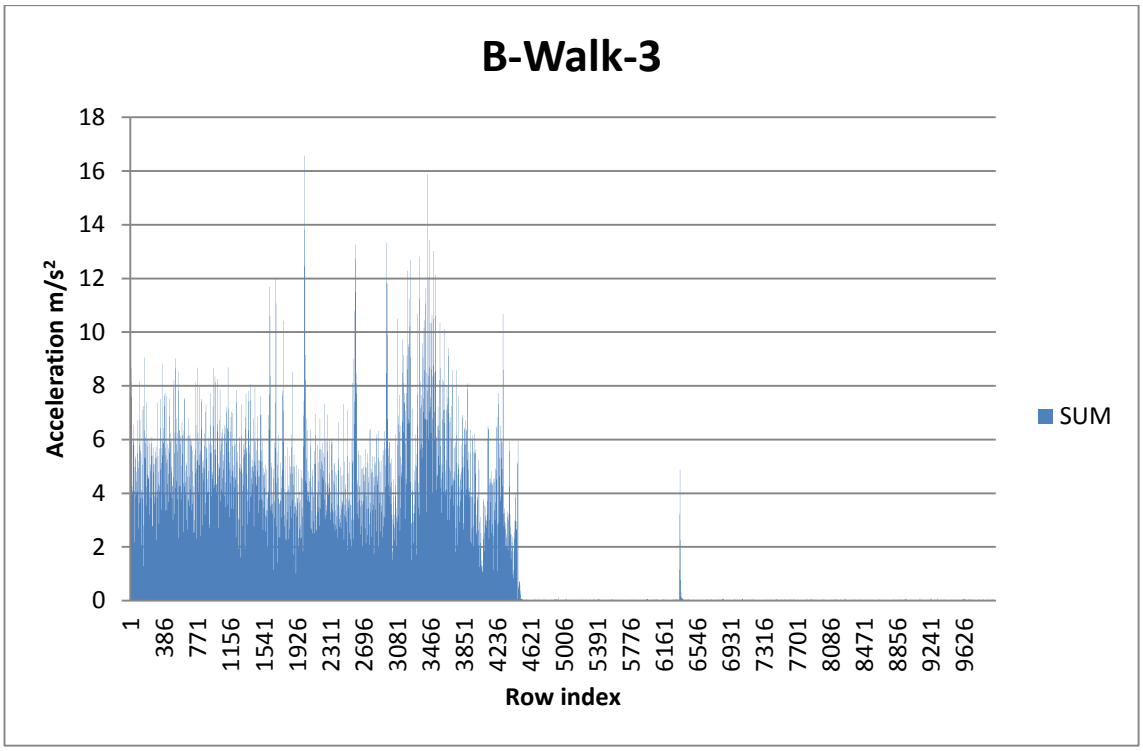

Figure 3: B-Walk-3 had fairly clear superfluous ambience at the end

B-Car-1 was fairly long (around 45 minutes) and interestingly had more intense accelerations towards the end. Although the B-Car-1 was not perfectly homogeneous from beginning till the end, there was no obvious reason for leaving parts of it out. It could for example be that first two thirds of the data was from low traffic highways and last third was from heavier traffic (Figure 4).

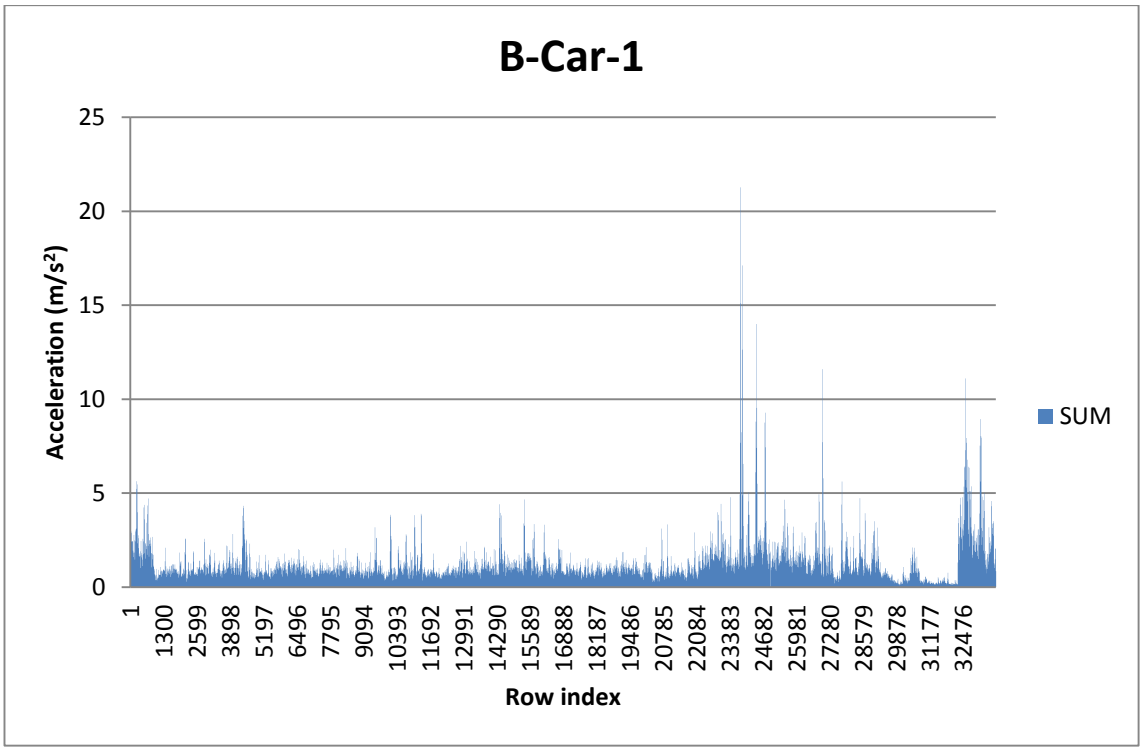

Figure 4: B-Car-1 contained more intense acceleration readings towards the end

3.4.3 Subject C

Subject C had only one very long (over 5 hours) bus data. The whole 5 hours was fairly homogenous with fairly low accelerations throughout the data which suggests it might really be one really long bus trip. Unfortunately this data seemed to have similar constant acceleration than that of subject A and had therefore be discarded from the analysis. As with 'A' it is likely that the subject C had an old version of the context logger.

3.4.4 Subject D

Subject D had recorded a great number of different transportation activities; in total 4 bus instances, 2 metro instances and 15 walk instances. Unfortunately it was very obvious that all of D data was corrupted in a similar way than that of subject C and A cases. See figure below (Figure 5).

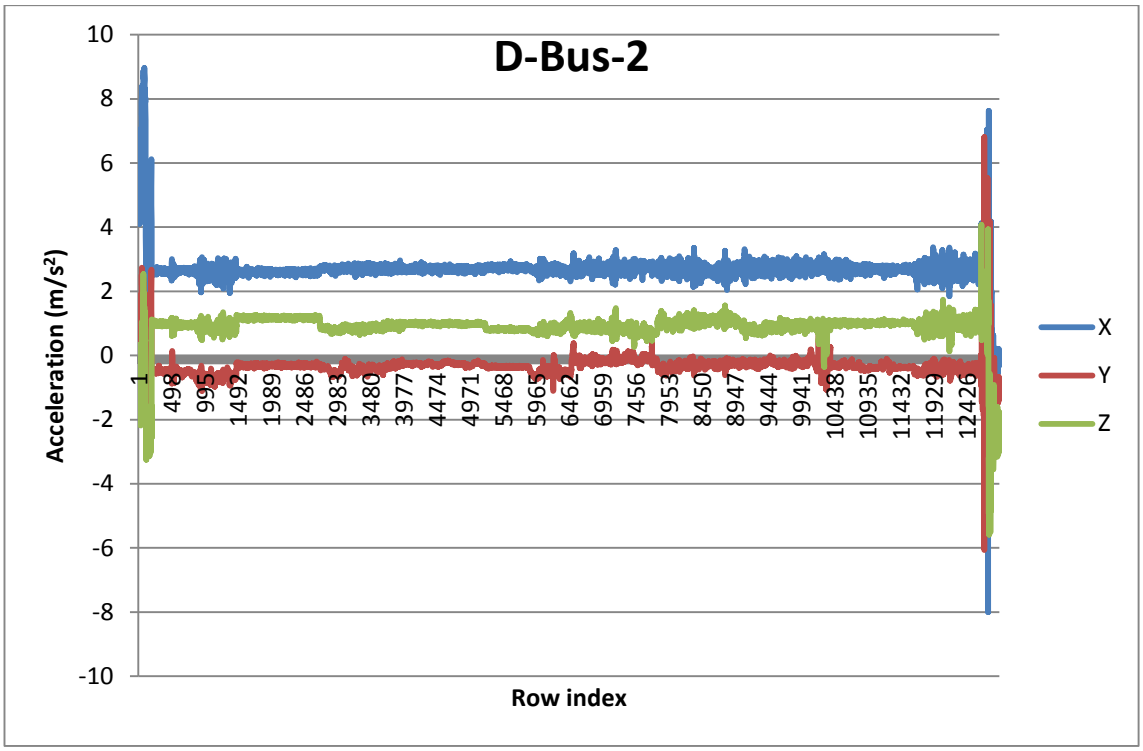

Figure 5: Subject D data had faulty constant acceleration readings

Although faulty, the data still nicely illustrates heavy accelerations at the beginning and end of data. These heavy accelerations were likely caused by the person using the device to start and stop recordings. Similar phenomenon was seen in many other instances too. Beginnings and ends of data sets were avoided with all subjects because of the very same reason.

3.4.5 Subject E

Subject E had a good number of instances from multiple activities: 4 bus, 9 walk, 2 car and even 1 train. Unfortunately the device E did not seem to provide as solid data as other devices. E had a lot more gaps in the middle of data which resulted in a situation where it was not possible to collect as many batches as from other subject's data. Some instances did not actually have any readings in them. However it was still possible to collect multiple batches from E-Bus-3, E-Walk-6 and E-Train-1. Gaps with the device E were likely a result of context logger software configured to save battery and not record all the time.

3.4.6 Subject F

Subject F had a great number of good quality data; in total 3 bus, 26 walk, 4 car and even 1 cycling, 2 taxi, 2 tram and 2 ski data. Cycling data was too short to collect any 500 reading batches from, but otherwise there were no evident reasons to discard any of the data. To avoid bias towards F not all of the data from this subject was used in final analysis.

Some of the data from F was a bit questionable while not being decisively corrupted. About the first half of the walk instances were fairly non homogenous. This probably should not result into direct discarding of all F-walk data. It however gives a hint that possibly not all the data within one instance is from uniform activity. This is not a problem in a situation where the person is indeed walking for the whole length of the recorded instance. Maybe he has only changed the walking rhythm. Non-homogenous data, however, can be seen as a problem if the person performs completely non-walking-related actions such as stands still in red lights or in a crowd. The problem is that there is no way to tell what they were really doing. Main guideline was to discard only obviously faulty data and in general to trust what the persons have logged.

F-Ski-2 is really interesting data (Figure 6). It takes place shortly after the much longer F-Ski-1, which seems fairly homogenous and credible data. F-Ski-2 in contrary has very unique acceleration pattern not seen in any other data. Its location in timeline and the pattern suggest that it might be for example one ski down the hill. Perhaps the test subject wanted to test out what kind of acceleration pattern this would result to. Skiing was not one of the key activities to be recognized, but some batches were still collected from F-Ski-1 to compare them with similar walk data. F-Ski-2 was not used in analysis as there would be no way to recognize it by using the approach used in this thesis and by using 10 second time frames.

While having similarities with the walk instances, F-Ski-1 on the other hand had more complex rhythm (see for example F-Ski-1-05 in Figure 7). If the task was trying to separate walk data from ski data – or perhaps different kinds of walk and ski styles – FFT and switching into frequency domain would probably come in handy.

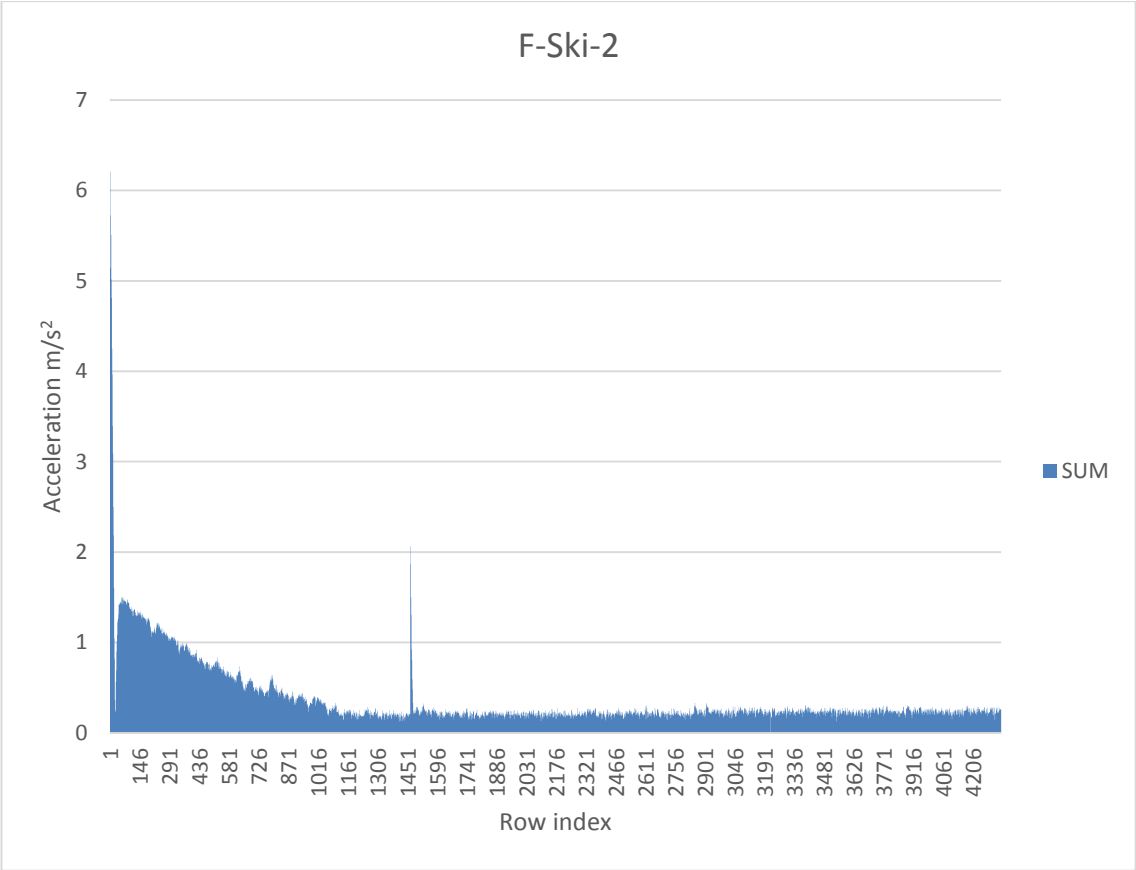

Figure 6: F-Ski-2 had really interesting shape. It might be faulty data or perhaps the test subject wanted to try what kind of acceleration pattern would be the result of some specific kind of skiing activity - perhaps skiing down the hill

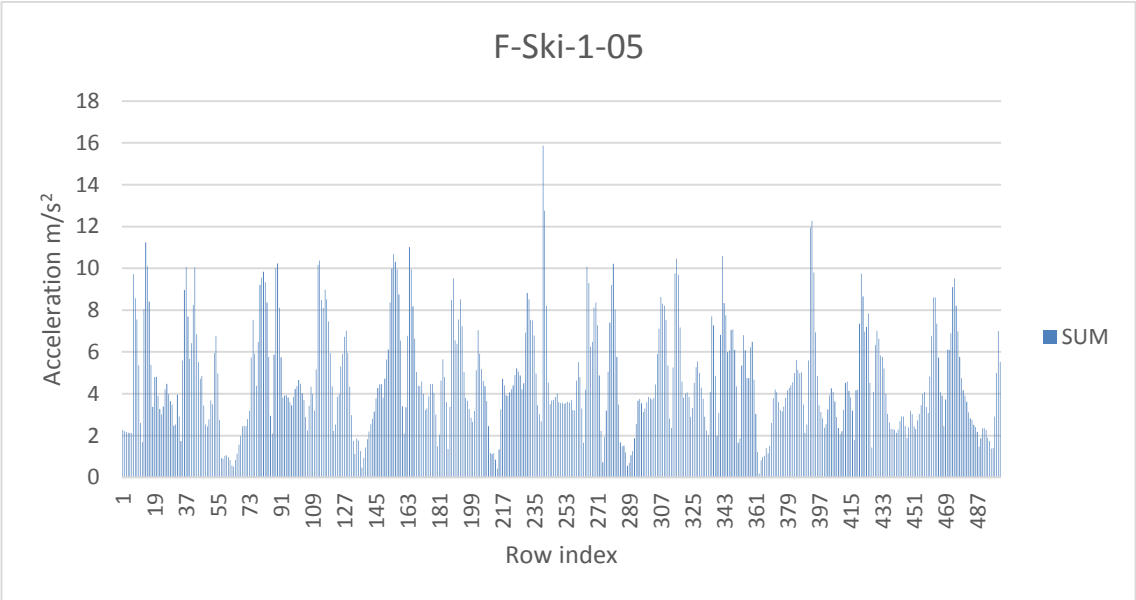

Figure 7: F-Ski-1 was somewhat similar with walk instances but with a bit more complex rhythm.

3.4.7 Subject G

Subject G had collected 9 instances of car data and 6 instances of walk data which was quite a lot when compared to other subjects. As with subject F not all available data was used in the analysis to avoid bias. Car data was chopped into 10 batches and walk data into 20 batches.

Device G seemed to have a unique property of recording couple of zero acceleration readings right after the start of tagged instances. This is merely a curiosity more than anything since as stated before beginnings and ends of instances were avoided in any case.

3.4.8 Subject H

Subject H had collected 1 instance of bus data and 4 instances of walk data. The collected data was in parts questionable because of non-homogeneity issues. Because of non-homogeneity some human decision making was needed to evaluate what should end up in analysis. Especially questionable were H-Walk-3 (Figure 8) and similar H-Walk-4 which were not used as it was really hard to say what parts the data were actual walk data. Clearly there were parts when the subject had been standing still. The non-homogeneity of H-walk-data suggests that he/she had collected the walk data in more relaxed conditions compared to other subjects. Perhaps H-walk-data included activities such as going out with a dog.

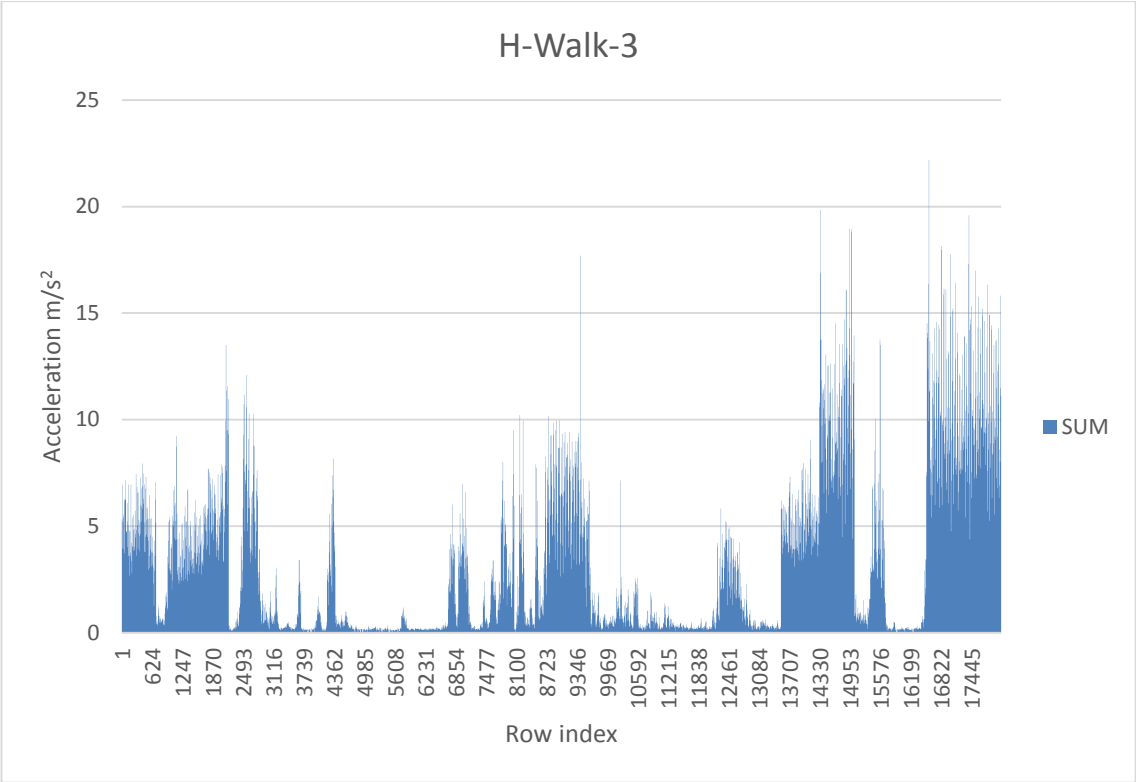

Figure 8: H-Walk-3 is an example of a non-homogeneous data where parts of the data are probably relevant but clearly there are parts which are not from "standard walking activity"

H-Bus-1 and H-Walk-1 were also somewhat non-homogeneous, but it was still possible to recognize main trend from them and collect the batches only from those parts.

H-Walk-2 was very homogeneous but also very short data and maximum amount of two batches was collected from that.

3.4.9 Subject I

Subject I had 2 bus and 2 walk data. All instances were credible, but not perfectly homogenous. As with all the subjects, only data that was obviously faulty was rejected and there was no reason why any major parts of I's data should be left out. All the data was chopped into batches for further analysis. We were aware – as with other subjects as well – that this kind of non-homogeneous data could, for example, originate from person using the device while recording data. But if there were no evident reason to reject something, we did not do so. Unless there were something like constant acceleration to certain directions, such as with 'A' and 'D', or sections of negligible acceleration while the person was walking, such as with H, everything was accepted to the analysis. The assumption was that the person had simply recorded the data in varying conditions. Perhaps part of

the bus data was while the bus was stopped and perhaps part of the walk data was from walking in a crowd.

Trouble when deciding which data should be included in analysis can be well demonstrated by looking at I-walk-1 which is presented In Figure 9. Purest walk data can be found in the middle of this instance whereas the beginning of the data was more non-homogeneous. After the row 4800 walking clearly ends and last acceleration spikes probably relate to picking up the device and ending the recording. Data from the beginning till the stoppage was chopped into I-Walk-1-01 - I-walk-1-09 although not all of it was homogeneous.

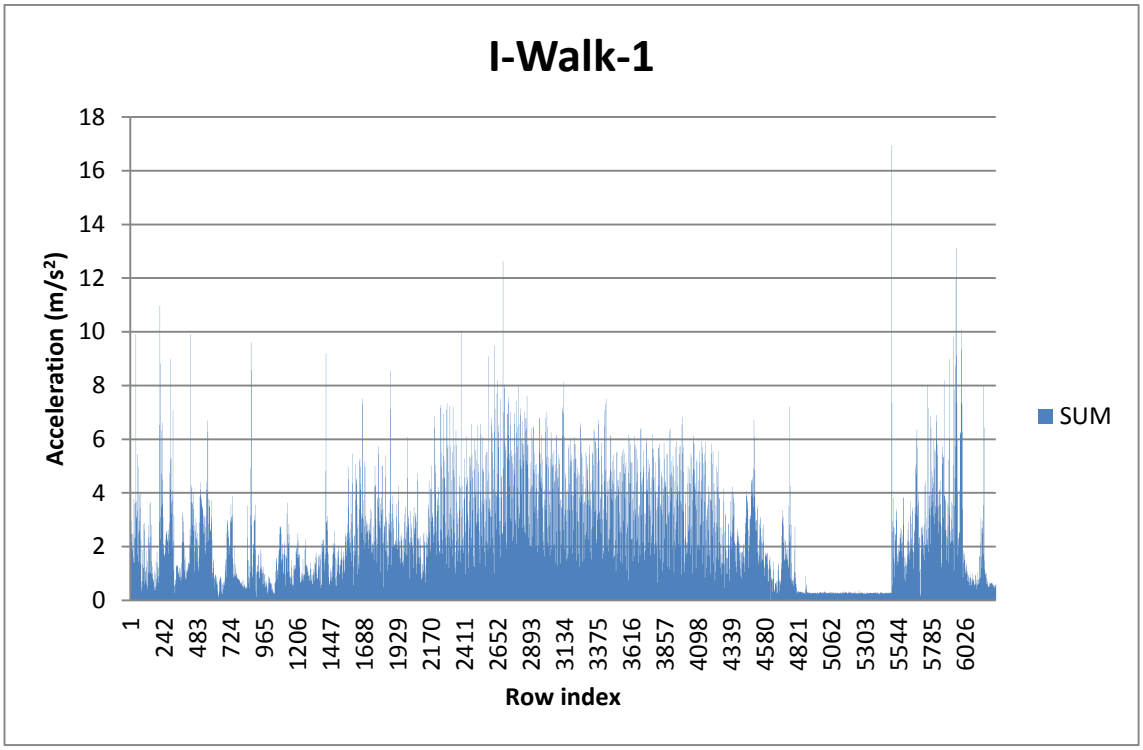

Figure 9: Roughly first 4800 readings of I-walk-1 data are probably actual walk data and accelerations at the very end represent activities needed to end the recording

Let's take a closer look to the collected batches. In the following two figures we present I-Walk-1-08 (Figure 10) and I-Walk-1-09 (Figure 11). These two batches very clearly demonstrate a rhythm typical to walking which was seen in most of the collected walk data. The latter of the two has typical walking rhythm in the beginning of the batch, but then fades into something more incoherent – similar to what the beginning of I-Walk-1 looked like. It looks as if the regular rhythm of the steps stops around row 160 and while the movement continues, it is not regular walking anymore.

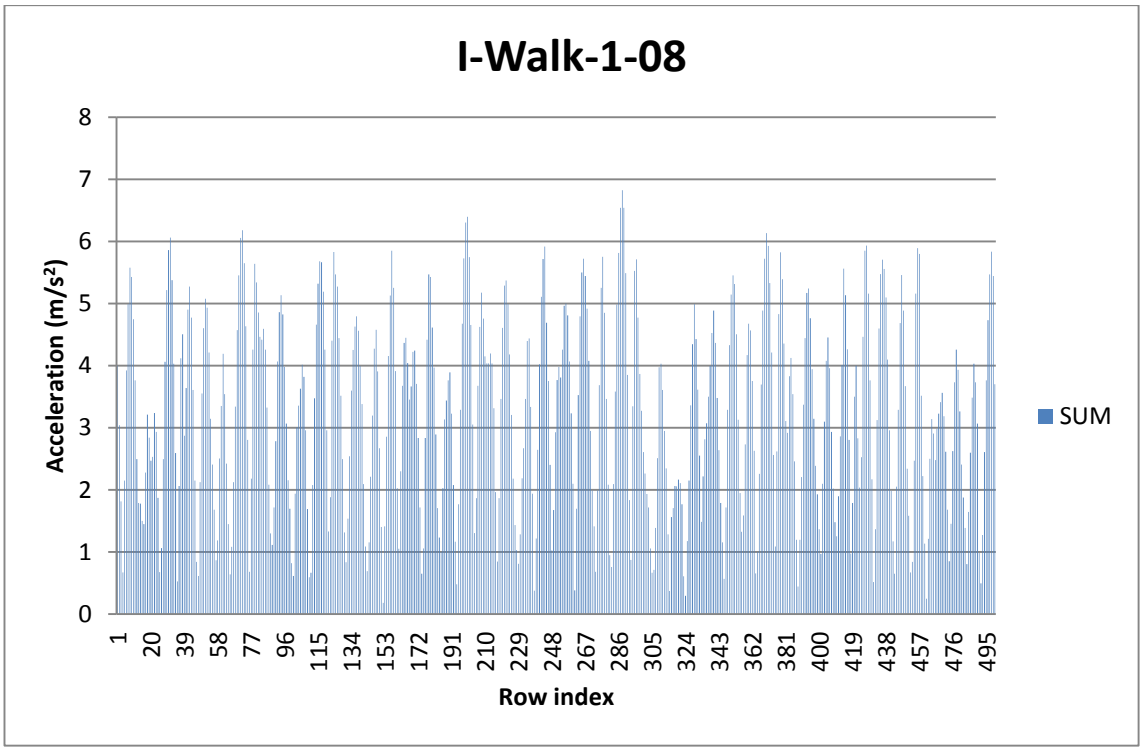

Figure 10: I-Walk-1-08 clearly demonstrates the rhythm typical for all walking data collected

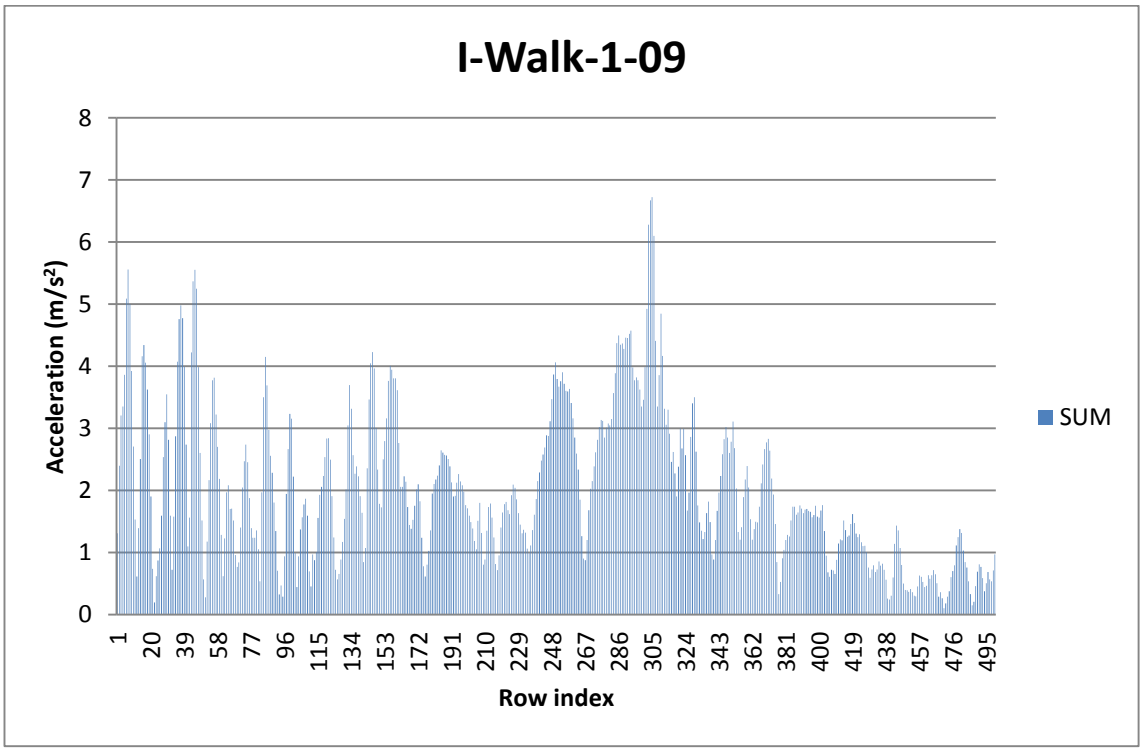

Figure 11: I-Walk-1-09 has typical walking data rhythm in the beginnning but transforms into something less coherent towards the end

3.4.10 Overview of the Clustered Data

For the final analysis data was picked from different subjects and from different instances as evenly as possible. From those subjects where less data was available more was used and from those who had collected plenty, not everything was used.

Main activities to be recognized were: walk, bus and car. Most of the collected batches thus originated from those instances. However since there was some data available also other transportation related activities some batches were also collected from: ski, tram, train and taxi instances. Batches used in final analysis are distributed as presented in Figure 12 (by test subject) and Figure 13 (by activity).

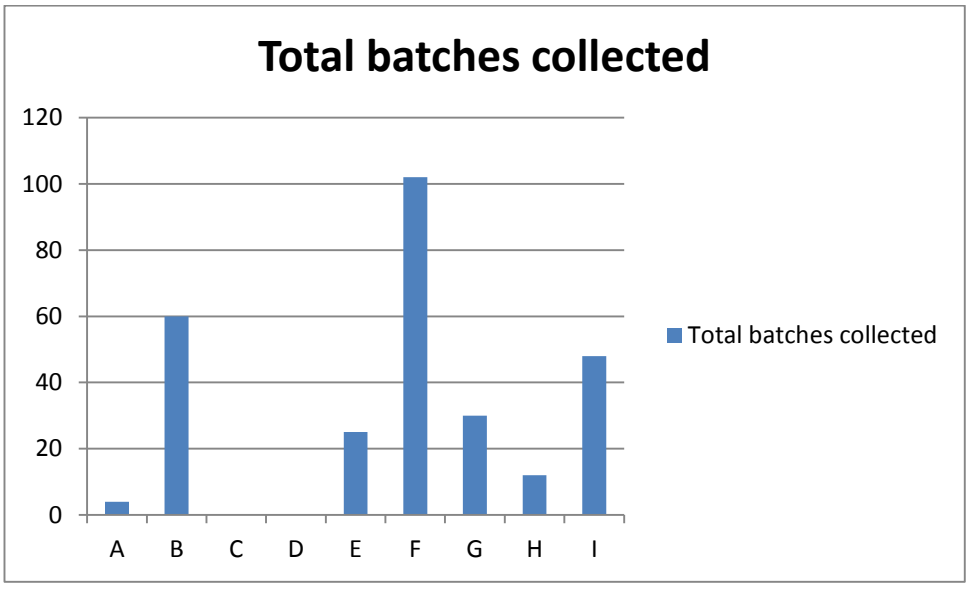

Figure 12: Final analysis data by subject

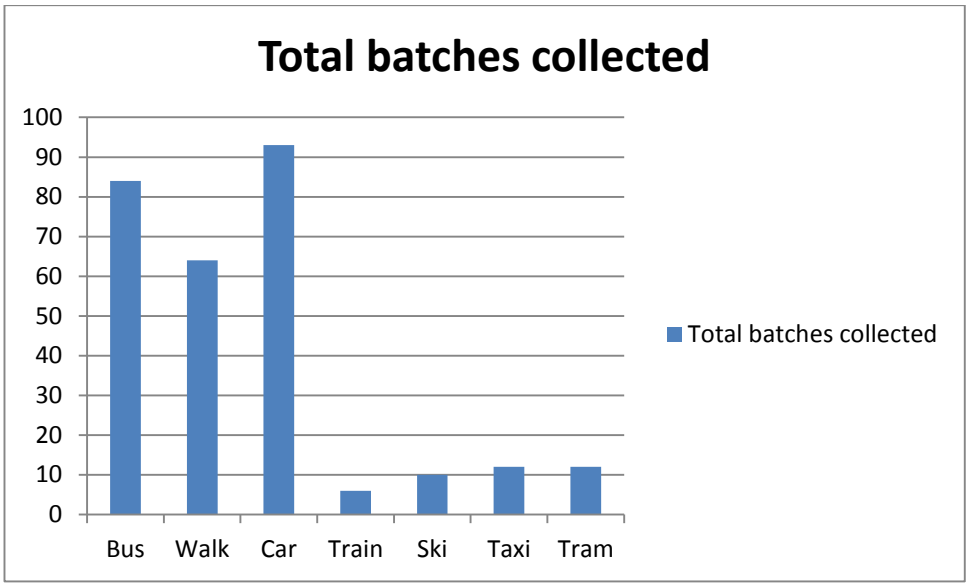

Figure 13: Final analysis data by activity

4 Results

The results section is organized under two main headings: “Optimal Clustering Method and Parameters” (4.1) and “Applying the K-means ” (4.2). In chapter 4.1 six different clustering methods were tested for partial data with different parameters. The aim was to find a method which would suit well for the type of data we had collected. In section 4.2 the best method was applied for full data. Clustering method cannot take pure accelerometer readings as an input so we calculated certain statistical variables to represent each of the 500-reading batches. Statistical variables used were: mean, median, maximum, minimum and standard deviation.

Partial data which was used in 4.1 consisted of 5 bus data from subject I (1-5), 5 bus data from H (6-10), 4 walk data from I (11-14) and 5 walk data from H (16-19). Distance method used was Euclidean distance and suitable cluster cutoff value was determined “on the fly” based on what the dendrogram looked like in each case.

4.1 Optimal Clustering Method and Parameters

4.1.1 Single Linkage (Nearest Neighbor)

With nearest neighbor clustering the optimal cutoff-value appeared to be little bit above 0,6. When for example 0,65 was used the dendrogram looked as presented in Figure 14.

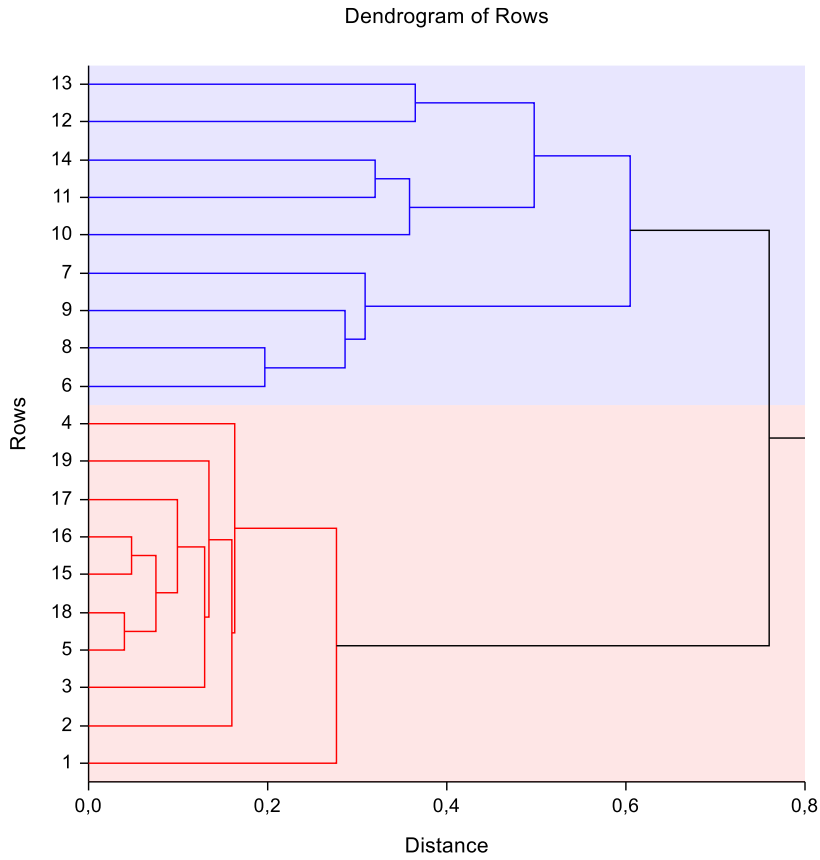

Figure 14: Nearest Neighbor clustering perfectly separated the walk data from bus data with cutoff value 0,65

The method seemed to cluster walk data and bus data from both subjects perfectly. It can also be noted that with suitable smaller cutoff value it would have been possible to cluster also the bus data from each subject into their own clusters (see dendrogram below), but the walk data was very solidly in its own cluster which also seemed to be a very solid general finding when experimenting with different kinds of data and with different kinds of methods.

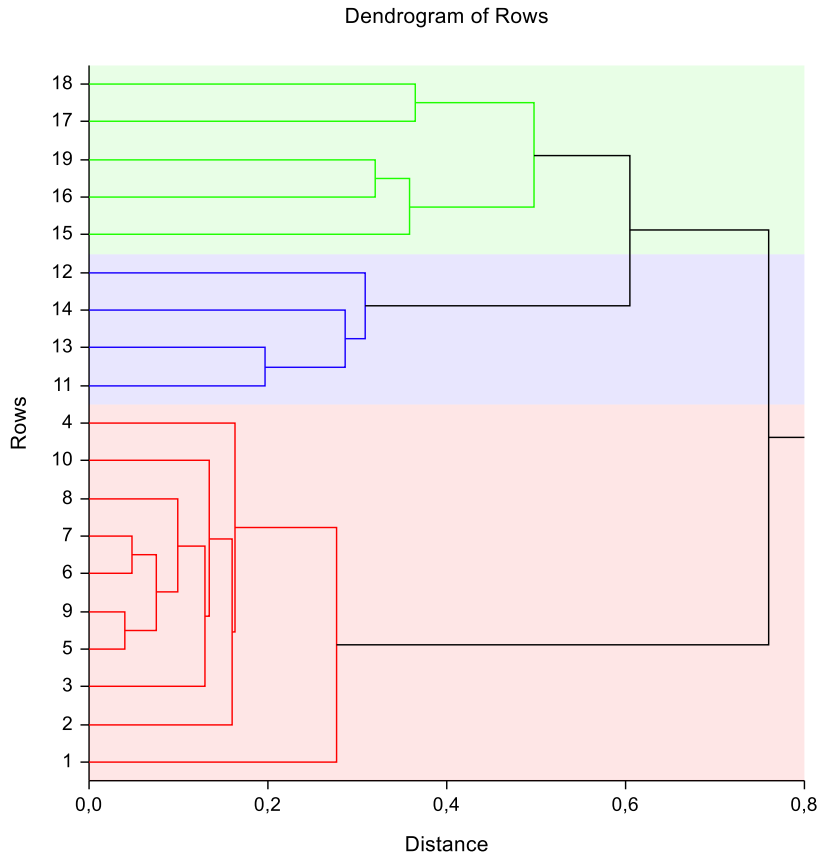

Figure 15: With smaller cutoff-value the nearest neighbor method separated walk data into its own cluster and bus data from two different test subjects into their own clusters.

As explained in chapter 3.3.1 Cophenetic Correlation and Delta-values can be used to assess the performance of clustering. The larger the cophenetic correlation the better and the smaller the delta values the better. However the values only tell a tale about the "mathematical performance" and do not consider which cluster the data points actually belong into in reality. That must be evaluated by human judgment. The values were as presented in Table 2.

Cophenetic Correlation	0,794
Delta (0,5)	1,103
Delta (1,0)	1,437

Table 2: Cophenetic Correlation and Delta values for Nearest Neighbor Method

4.1.2 Complete Linkage (Furthest neighbor)

With the furthest neighbor method there was not any cutoff value which would correctly cluster the data. Bus data from person I formed its own cluster but the problem was that person I “bus cluster” tended to be closer to I & H “walk cluster” than person H “bus cluster”. With cutoff of 1,2 bus data are clustered into their own clusters per person and walk datas from both persons form a single walk cluster. When cutoff of 1,5 was used the bus data from I formed the cluster with all walk data and subject H bus data formed its own cluster. See the following two figures (Figure 16) and (Figure 17).

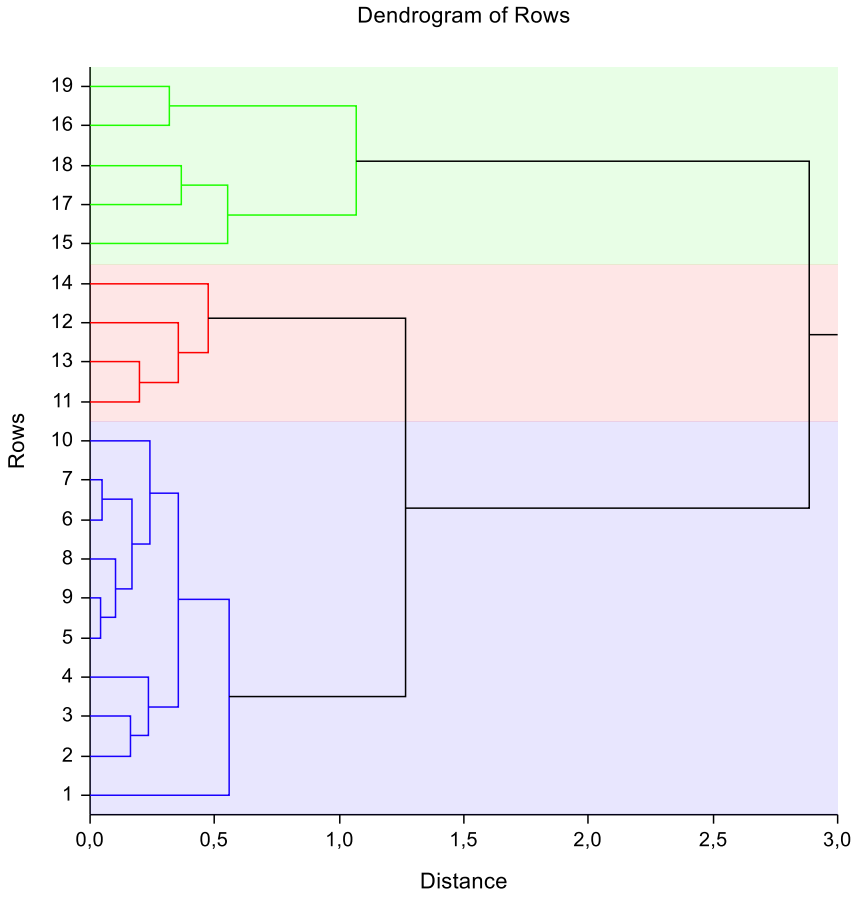

Figure 16: Problem with Furthest Neighbor Method was that significant driver in forming the clusters was test subject and not the mode of transportation.

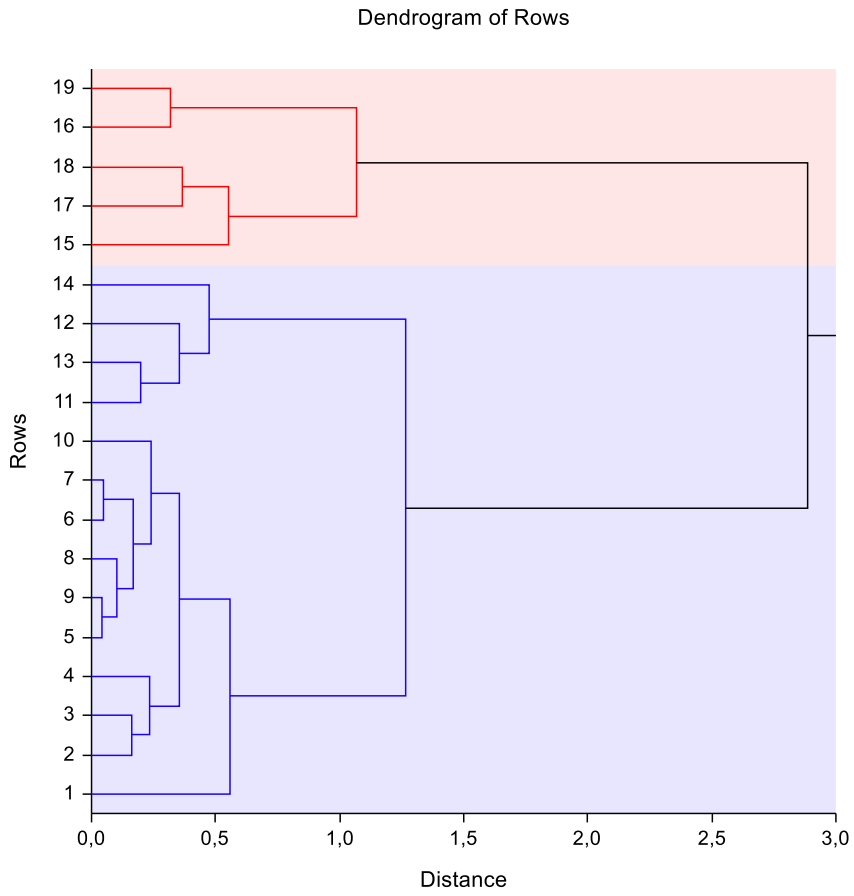

Figure 17: With smaller cutoff value it is possible to force formation of two clusters but then subject I bus data gets mixed with the "walk cluster"

Cophenetic correlation and delta values for furthest neighbor method are presented in Table 3.

Cophenetic Correlation	0,868
Delta (0,5)	0,290
Delta (1,0)	0,376

Table 3: Cophenetic Correlation and Delta values for Furthest Neighbor Method

4.1.3 Centroid Method

With centroid method it was impossible to find any cutoff value which would lead into one bus data cluster. A fairly good result could be achieved with cutoff value 1,0. Then the walk data formed its own cluster and the bus data from I and from H formed their own clusters. Now the subject's H bus cluster was not merged to one as with was the case with previous methods. In addition to that, one specific walk data (row 1) from person I also

tended to slip out from the “walk cluster”, but was still classified as walk even with quite small cutoff values. See the image below:

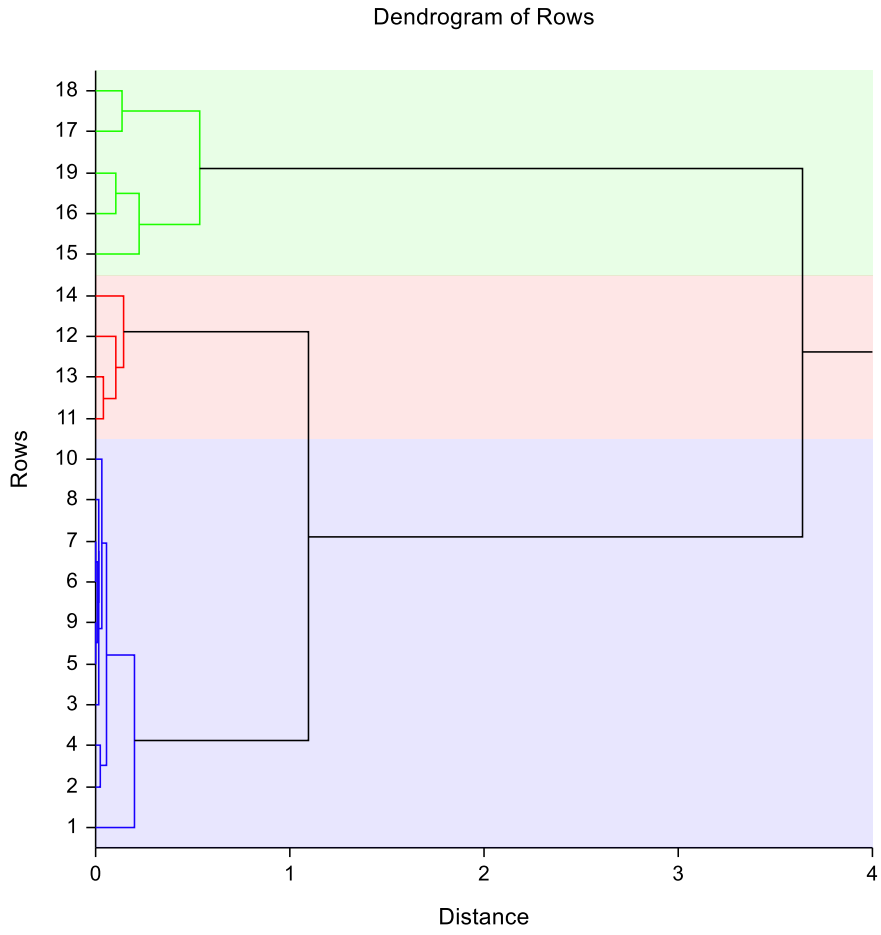

Figure 18: Centroid Method had same problem with Furthest Neighbor Method - a significant driver in forming the clusters was test subject and not the mode of transportation.

Cophenetic correlation and delta values for centroid method are presented in Table 4.

Cophenetic Correlation	0,779
Delta (0,5)	0,464
Delta (1,0)	0,612

Table 4: Cophenetic Correlation and Delta values for Centroid Method

4.1.4 Group Average Method

Using group average provided results which were very similar to those of furthest neighbor method, but were slightly better in the light of cophenetic correlation and delta values. Below is the dendrogram with cutoff value 1,0 which forms three clusters: All walk data, I bus data, and H bus data.

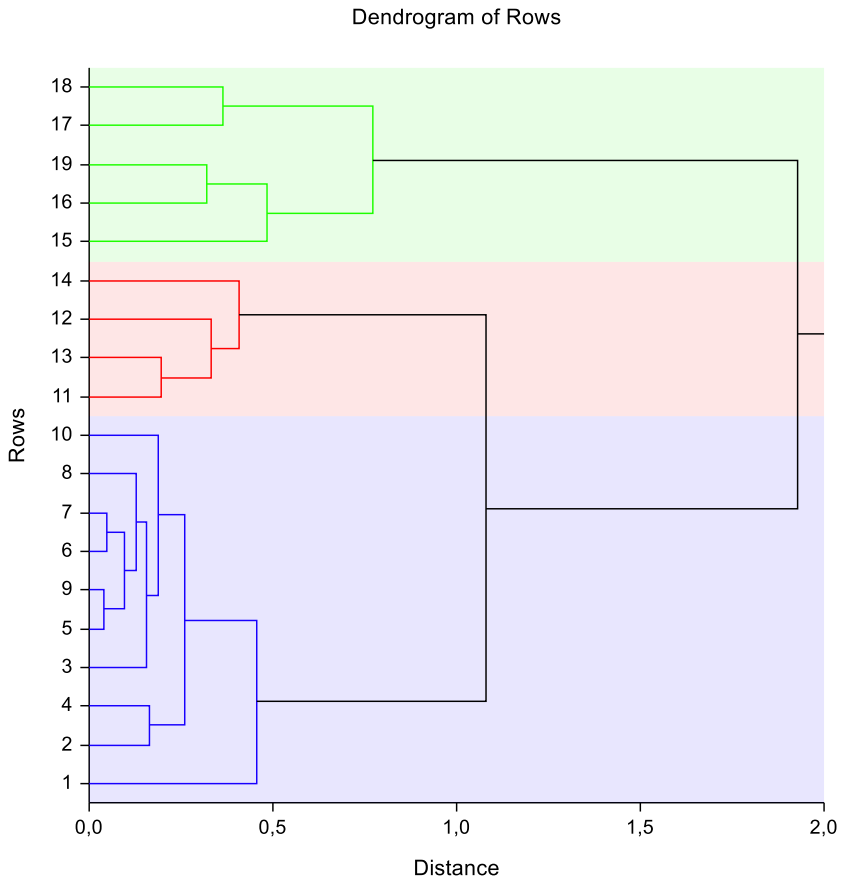

Figure 19: Test subject instead of transportation mode was a significant driver in forming the clusters also in the case of Group Average Method

Cophenetic correlation and Delta values for group average method are in Table 5.

Cophenetic Correlation	0,875
Delta (0,5)	0,213
Delta (1,0)	0,291

Table 5: Cophenetic Correlation and Delta values for Group Average Method

4.1.5 Ward's Minimum Average Method

Ward's minimum average method formed really clear bus and walk clusters, but was not as good in the light of cophenetic correlation for which the values bigger than 0,75 are generally seen as good. Cutoff value 8,0 provides the following dendrogram (Figure 20) with bus and walk data in their own clusters.

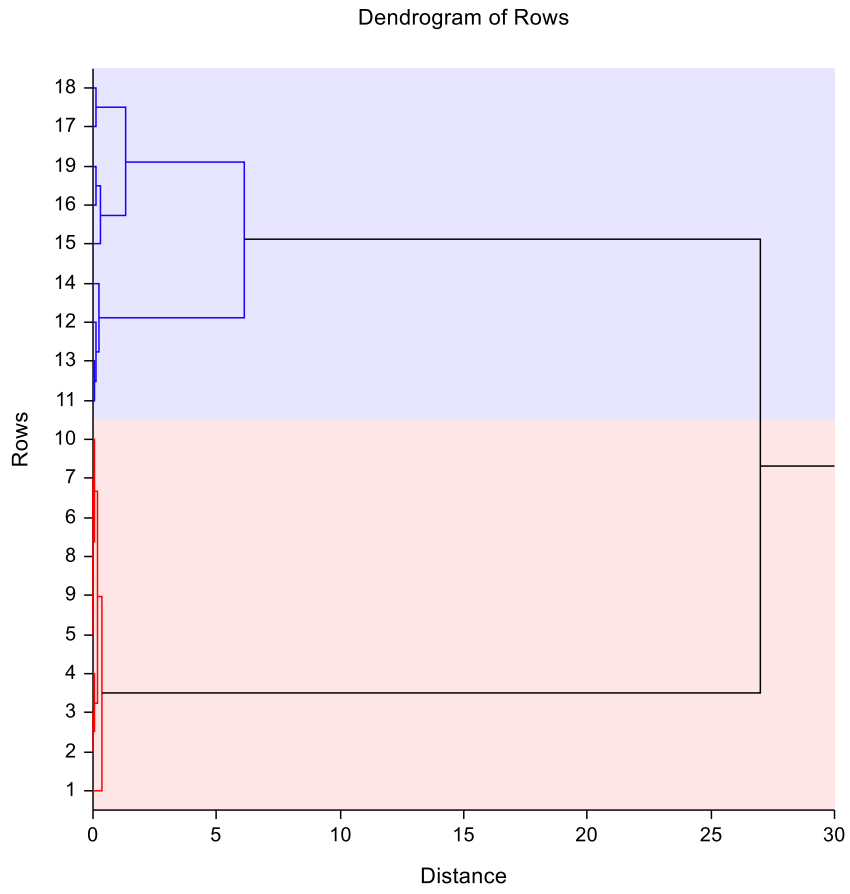

Figure 20: Ward's Minimum Average Method perfectly separated the walk data from bus data - a similar result than that of Nearest Neighbor method.

Cophenetic correlation and Delta values for Ward's minimum average method are presented in Table 6.

Cophenetic Correlation	0,640
Delta (0,5)	0,867
Delta (1,0)	0,879

Table 6: Cophenetic Correlation and Delta values for Ward's Minimum Average method

4.1.6 Overview of the Hierarchical Method Results

The nearest neighbor method seemed to perform best of the hierarchical methods. It separated the walk data perfectly from bus data and could actually also separate the two different test subjects' bus data from each other with suitable cutoff value. However it must be noted that Delta values of nearest neighbor method were the largest of the five,

which suggests its mathematical performance could have been better. Cophenetic correlation value was however above 0,75 which can be usually seen as good.

Ward's minimum average method performed very similarly but cophenetic correlation was below 0,75. Furthest neighbor method, centroid method and group average method all had trouble clustering the batches by activity and they tended to rather cluster it by subject.

The comparision of different methods (Table 7) shows that the nearest neighbor method performed best followed closely by Ward's minimum average method.

	Cophenetic Correlation	Delta (0,5)	Delta (1,0)	Clustering Success
Nearest Neighbor	0,794	1,103	1,437	Perfect
Furthest Neighbor	0,868	0,290	0,376	Major driver was subject and not activity
Centroid Method	0,779	0,464	0,612	Major driver was subject and not activity
Group Average Method	0,875	0,213	0,291	Major driver was subject and not activity
Ward's Minimum Average Method	0,640	0,867	0,879	Perfect

Table 7: Comparision of different hierarchical methods

4.1.7 K-means

K-means differs from all previous methods by being non-hierarchical. Knowledge about the amount of clusters can now be assumed before the analysis. Thus the result plots are presented in different format and there are no more dendrograms. Because of this the K-means results are not fully comparable to the previously used hierarchical methods.

With test runs, the number of different transport methods was two (bus & walk) so the amount of reported clusters was set to two hoping that analysis would separate walk batches from bus batches. Random starts was set to 3, Max iterations was set to 25 and

Percent missing was set to 50. Minimum and maximum amount of clusters during the analysis was set to 2 and 5 respectively. Different options were experimented with but adjusting the parameters did not change the results considerably. The best result was found by using these values.

Clustering with k-means worked really well. In the two clusters which were formed there was only one wrongly placed data point: row 12 – I-Walk-2-02 which ended up into bus data cluster. See for example Average vs. Std. deviation plot below (Figure 21).

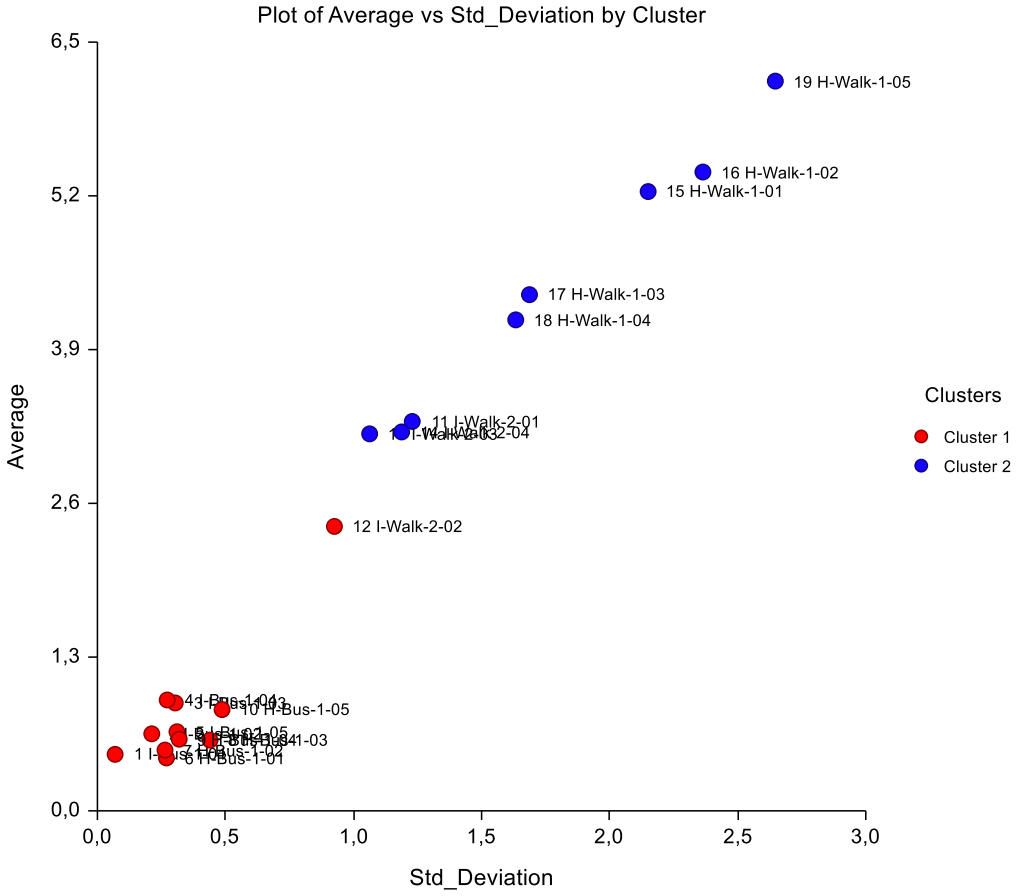

Figure 21: K-means clustering separated walk batches well from the bus batches.

4.2 Applying the K-means Method to All Data

Non-hierarchical K-means method seemed to work particularly well for clustering the kind of data we had collected in the 19 batch test runs while nearest neighbor method seemed to be the best hierarchical method.

The superiority of K-means became clear when it was compared to hierarchical methods with all collected batches. When the number of batches got high, it became impossible to find suitable cutoff value for getting any kind of satisfactory results with hierarchical methods.

Regardless the cutoff value the analysis always seemed to give few really small, internally similar clusters and then one really big cluster containing almost all other batches. For example, applying the best combination of test runs – Nearest Neighbor method with a cutoff value 0,65 – results into three clusters:

1) B-Walk-2-02 & B-Walk-2-04
2) H-Walk-2-01 & H-Walk-2-02
3) All the rest of batches

Dendrogram with all available batches becomes fairly unreadable but is still presented in Figure 22 (Nearest Neighbor with cutoff value 0,65).

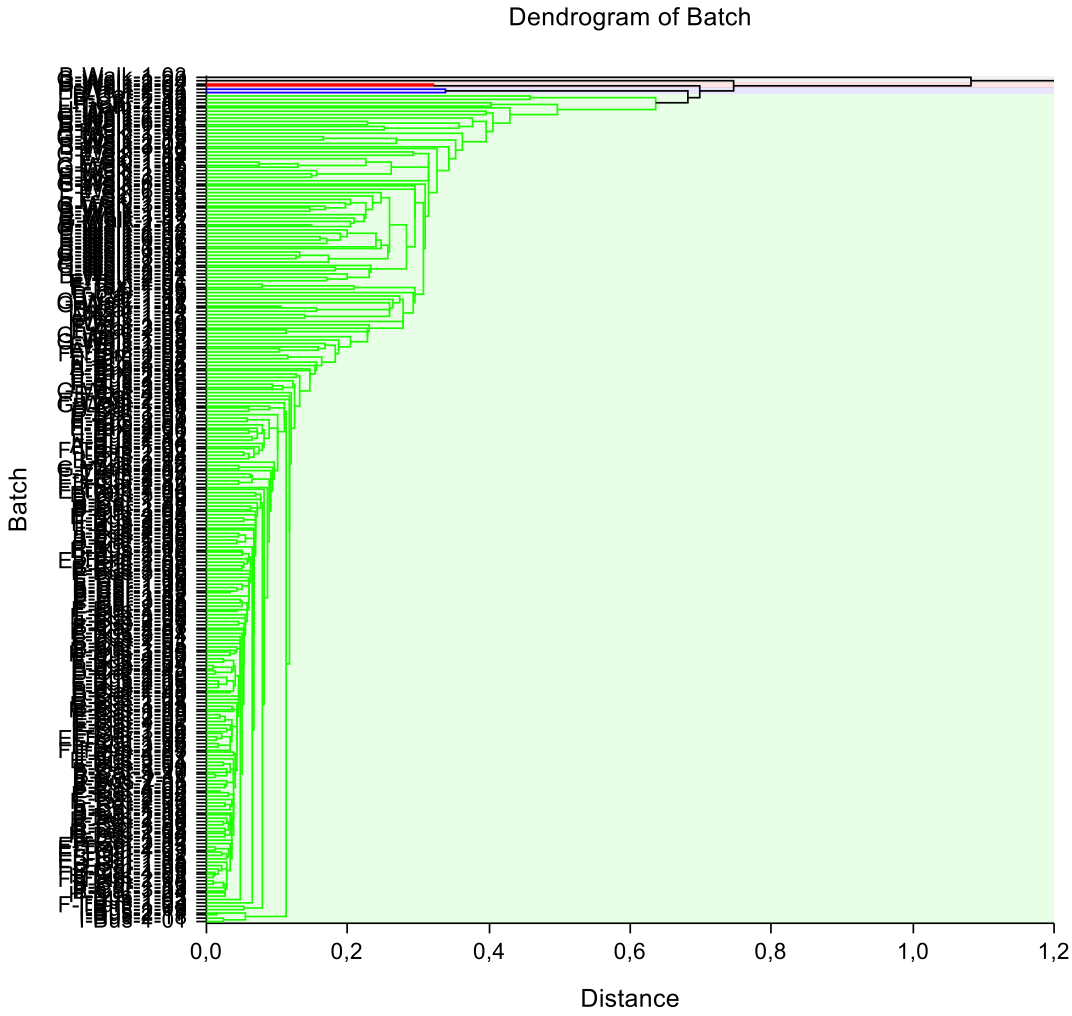

Figure 22: The nearest neighbor (which performed best of hierarchical methods in test runs) did not find a cutoff value for producing reasonable results.

K-means seemed to provide sensible results even when the number of batches got higher. Note that we now had also few batches of ski, train, taxi and tram available and different amount of reported clusters were experimented with. For example using reported clusters of 4 resulted into clusters presented in Table 8.

Table 8: Clusters formed by applying K-means to all data with 4 reported clusters

Label	Car	Bus	Walk	Ski	Tram	Train	Taxi	Total
“Walk/Ski”	5	3	34	3	1	-	-	46
“Motor”	2	6	-	-	1	-	5	14
“Motor”	76	75	5	-	10	6	7	179
“Walk/Ski”	-	-	25	7	-	-	-	32
Total	83	84	64	10	12	6	12	271

While clustering algorithm formed the groups, labeling them relied on human interpretation. Clusters 1 and 4 seemed to be walk and ski clusters with cluster 1 having some misplaced batches. Clusters 2 and 3 seemed to be motorized transportation clusters with cluster 3 some misplaced walk batches.

In total 92,0% of all walk data were assigned to either cluster 1 or cluster 4. In total 95,2% of all car and bus data were assigned to either cluster 2 or cluster 3. On the other hand, nothing decisive could be said about separation of car and bus data. They seemed to be very much alike.

The resulting graph can be presented by any of the statistical values used to represent the batches. In the figures below results are presented in a standard deviation vs. average plot (Figure 23) and min vs. max plots (Figure 24) which both present the separate clusters fairly well.

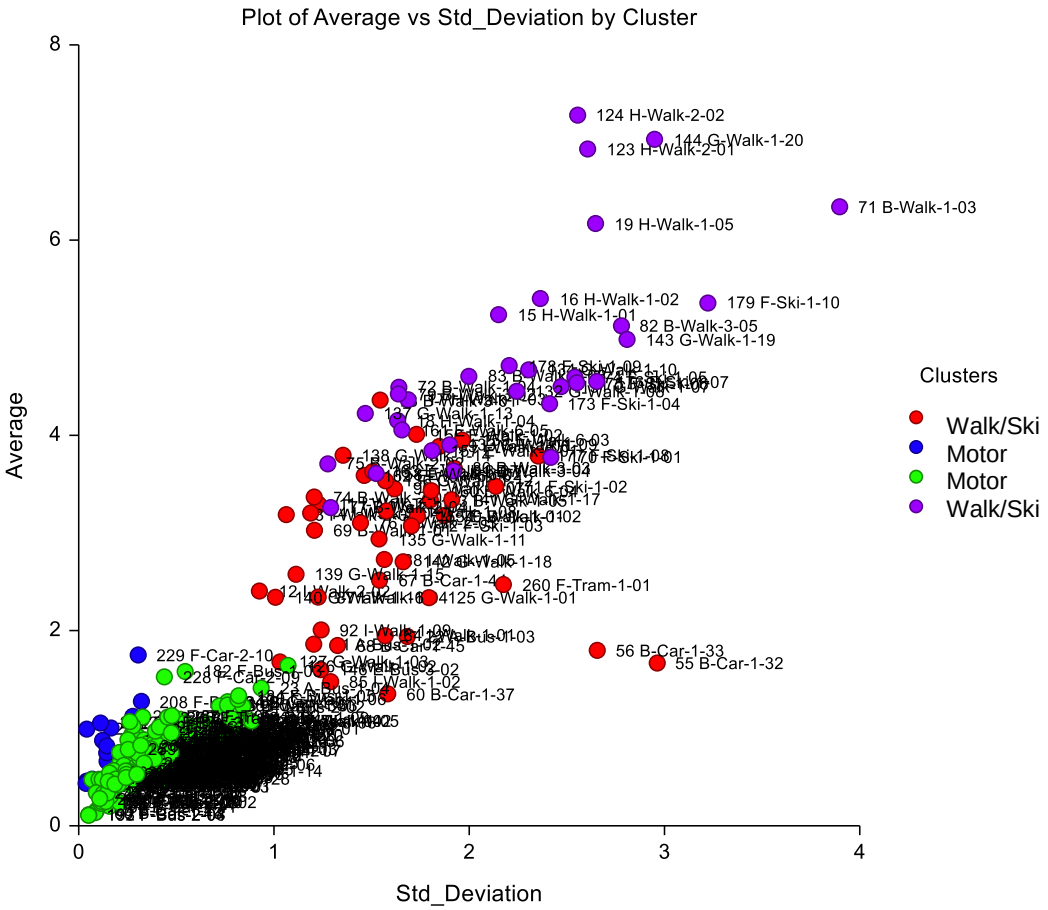

Figure 23: K-means clusters presented in a standard deviation vs. average graph

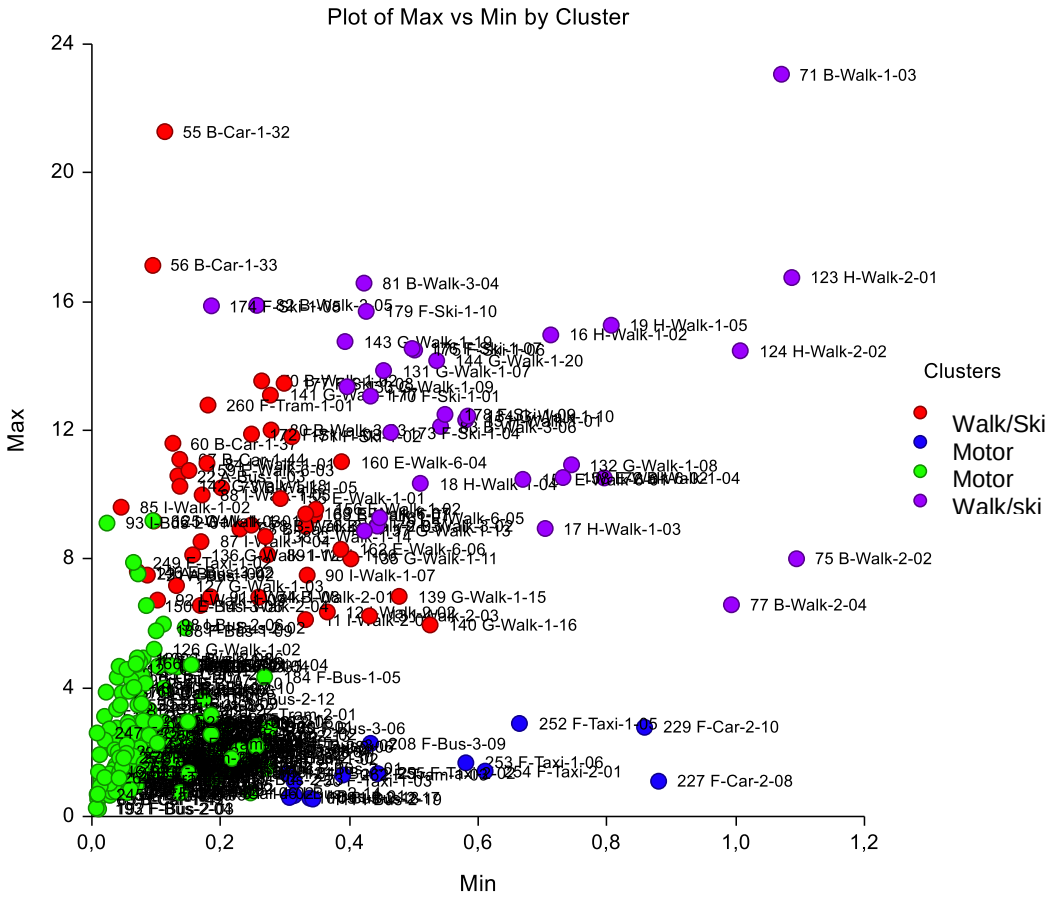

Figure 24: K-means clusters presented in min vs. max graph

When k-means was applied to only walk, car and bus data, the result in Table 9 were obtained. Result was similar to what we got with all available data but cluster 2 was not as clearly "motor" cluster anymore.

Label	Car	Bus	Walk	Total
"Walk"	3	-	32	35
~"Motor"	5	10	11	26
"Motor"	75	74	2	151
"Walk"	-	-	19	32
Total	83	84	64	231

Table 9: When only car, bus and walk data was analyzed, two walk-clusters, one motor-cluster and one mixed cluster with a motor bias were formed.

When the number of reported clusters was dropped to three – we had three different modes of transportation after all – the result was one fairly clean "walk" cluster, one fairly clean "motor" cluster and one "motor" cluster with some walk data. This was to be expected, considering the other runs.

Walk data always seemed to separate to its own cluster(s) but bus and car data were mixed. So what if we assumed that by using the chosen method and with the given data it is simply not possible to separate bus and car data? What if we set the reported clusters into just two – would one of them be "walk" and the other "motor"? Results are presented in the table below (Table 10).

Label	Car	Bus	Walk	Total
"Motor"	80	83	8	171
"Walk"	3	1	56	60
Total	83	84	64	231

Table 10: When reported clusters is set to two one of them is very clearly walk cluster and the other fairly clearly "motor" cluster

With two reported clusters one of them is very clearly (93,3%) walk cluster and other fairly clearly (95,3%) motor cluster while 87,5% of walk batches ends up into walk cluster and 97,6% of motor batches end up into motor cluster.

These were fairly good percentages, given that we allowed the test subjects to collect the data freely and were not very selective when picking the data for the analysis. Only obviously faulty data was discarded. Wrongly clustered walk batches possibly originated from situations where the subject has not been walking but instead stood, for example, in red lights. It is easy to believe that a person standing relatively still would produce similar acceleration profile to that of a person sitting in a car on straight road – at least now when relatively simple statistical numbers were used to describe the batches.

5 Discussion

5.1 Conclusion

The main objective of this thesis was to create a multivariate model which would separate different transportation modes into their own clusters based on accelerometer data collected from cell phones carried by test subjects. The collected accelerometer data was cut down to 500-reading *batches* which were then described with batch-specific statistical values: Mean, median, minimum, maximum and standard deviation. Various clustering algorithms were applied to the batches which were represented by statistical values listed above.

A non-hierarchical K-means method was especially well suited for our case. K-means performed well in the test runs and it also needs less human decision making than hierarchical methods. When applying K-means to all collected batches, batches originating from walking activity clearly formed its own cluster(s) and batches originating from motorized vehicles (bus, car) formed its own cluster(s). However, it was not possible to separate car data from bus data by using the selected approach.

It was also found out that using the data originating from loosely defined data collecting framework (i.e. letting the test subjects carry and use the data collecting devices just like they would do in everyday life) is problematic from the activity recognition perspective. Test subjects could well have their own ideas about what walking actually is and what is travelling by bus/car. For one person only uninterrupted long walk outside is walking while the other could tag any “pedestrian activity” as walking. Data tagged as walking could therefore contain e.g. window shopping or standing still on the red light.

Similarly, travelling by bus or car could contain data originating from various conditions. Some of the data can be from heavy traffic and some from a highway with little traffic. Some of the data was likely collected while the person was actually using the phone which can create different acceleration profile compared to having the phone in their pocket etc.

Regardless of the loose setting, walk data seemed to be so fundamentally different that it could be easily separated from car and bus data. This is somewhat intuitive result. Even with simple statistical numbers describing the collected batches, much higher accelerations with much higher variance are associated with walking than with sitting in a car. Accelerations related to driving a car are relatively low and relatively long-lasting with low variance, for example when the road is curving. If the person, on the other hand, is walking, every step while the phone is in their pocket can probably create accelerations that only rally driver could match.

5.2 Reliability and Validity of the Research

The data was collected from the nine different test subjects as in previous studies. However, it was unfortunate that a lot of the collected data had to be discarded before the actual analysis and from two test subjects all data was corrupted.

The nature of the study was experimental and exploratory. At first we only had vision about collecting cell phone accelerometer data with Contextlogger3 and then, by using multivariate methods we would recognize the modes of transportation. Other than that we had no clear vision about the final setup. We looked for hints from the literature and experimented with our own test data which we collected on-the-fly while developing the data collecting setup at the same time.

Exploring with test data was fun and exciting, but once the decision needed to be made on how to collect the real data, it was like closing a door with no certainty on whether we have made the right choices. Once the real data had been collected and processed for NCSS, it would have been *very* burdensome to go back in square one if that would have been needed.

Various decisions needed to be made and it felt difficult to start narrowing down degrees of freedom one by one while the suitability of the selected approach was still uncertain. For example the following decisions needed to be made:

- What type of sensors are used: Data can be collected by using e.g. cell phones and tablets, but also by designated accelerometer or other sensors.
- How the sensors are placed: Test subject can be instructed to carry the device in some specific way, such as in pocket or in backpack, or they can be allowed to carry the device freely.
- How well the subjects are trained for data collecting and how strictly the activities are defined: Test subjects can be allowed to act freely or they can be instructed to, for example, collect walking data only when they are not in a crowd which might disrupt their walking activity. Similarly for a car activity: It can be defined on what kind of roads they can collect the car data, are they allowed to collect data in traffic etc. In this study we didn't have any restrictions for the subjects. They simply performed their daily routines and used the context logger to log what they did.
- What methods are used for recognition: In this study we used multivariate methods, but there are also other possibilities, such as Hidden Markov Model or other machine learning models.
- How long are the analyzed datasets: 1 second, 10 seconds, a minute, or something else?
- What statistical numbers should be the inputs for clustering: If clustering method is used, it must be decided how the datasets are presented – clustering cannot take in pure accelerometer readings. We used simple statistical numbers: min, max, median, average, and standard deviation to represent the batches.
- Should there be some more advanced preprocessing of the data before clustering: We experimented a little with FFT, which did not provide much more information about the data. In our case only walk and ski data seemed to have specific rhythm.

Different combinations of these options lead to different results. There are combinations which provide very good results and some combinations do not provide any meaningful results. There were so many options that it was impossible to try everything with everything. We experimented and made decisions one by one based on what seemed to work.

If something had been chosen differently in early phases of the study, rest of the decisions could have taken very different route.

Based on the resources and competences available we decided to use cell phones for collecting the data and multivariate methods to later classify it. Same data was used also in other studies, and thus we allowed the subjects to carry the devices as casually as they would in everyday life. Experimenting with the setup suggested that the interesting phenomenon happened within the seconds' time frame so we decided to collect data at high intensity (50 readings/second) and use rather short time frame for batches (10 seconds). To represent the collected batches we used basic statistical numbers (min, max, median, average and standard deviation) which seemed to provide satisfactory results with our data. There could be some more elegant options available if, for example, the data collecting setup also was more strictly defined. We experimented with FFT, but its benefits would likely come out only when data is collected through a strictly defined setup. There are various clustering options to choose from; for us K-means seemed to provide the best results.

Better recognition results would probably have been observed had the data collecting framework been set up more strictly. If the aim had been to create as good of a model as possible it would be crucial to very strictly define what kind of data we want from different modes of transportation – where do we want the "center of a cluster" to be.

5.3 Suggestions for Future Research

We only used accelerometer data while the devices could have provided more. Especially GPS-data could be very helpful by providing an easy way to use velocity to enhance clustering accuracy. In theory, it is possible to get velocity also without GPS by integrating the acceleration data, but it was not possible in practice, because we would need to know instance-specific initial velocity v_0. Even if we could assume it zero, technical limitations, in the form of gaps in the data, would still be a problem. Calculating velocity from acceleration data would also be very sensitive to systematic errors. Even slight acceleration bias to one direction would eventually accumulate into significant (and faulty) velocity to that direction.

The collected data was also used in other studies so large part of total available data was actually not from transportational activities. The data was collected during a time frame of few days (there was slight person to person variance) while the test subjects performed normal daily activities and chores. There were some cases where a subject had forgot to mark the beginning or the end of an activity. We eliminated these cases through human judgement before clustering. We had not restricted the use of devices in any way so it is likely that some of the data was collected while the device was in normal use. This obviously would lead different acceleration profile compared to having the device in the pocket. A loose setup was acceptable, and to a degree even desired, because one of the study objectives was to identify challenges with real unknown data from real everyday life.

If the aim were to create as accurate recognition as possible, the data should be collected in shorter time frames and under stricter rules. Different modes of transportations and their characteristics should first be defined: In which proportions we want data from traffic and from open road? Should "car in traffic" and "car on open road" actually be different clusters? What is walking inside of a bus? Is the person allowed to use the phone while collecting the data? After the archetype for each transportation mode is defined – once "the center of each cluster" is defined – the data should be collected in sharp, systematic and activity-oriented stints.

We focused on comparing the performance of different clustering algorithms, but the focus could have been also on comparing other aspects. We could have chosen K-means at an early phase and focus on comparing different inputs for the clustering. We used mean, median, maximum, minimum and standard deviation to represent batches, but perhaps there are better and more sophisticated options available. For example, by analyzing the derivatives it would be possible to say how many times acceleration turns into deceleration within one batch. Based on what most of the walk data looked like, this could actually turn a cell phone into moderately good pedometer.

References

Aharony, N., Pan, W., Ip, C., Khayal, I., Pentland, A., 2011. The Social fMRI: Measuring, Understanding, and Designing Social Mechanisms in the Real World, in: Proceedings of the 13th International Conference on Ubiquitous Computing, UbiComp '11. ACM, New York, NY, USA, pp. 445–454.

Atzori, L., Iera, A., Morabito, G., 2010. The Internet of Things: A survey. Computing Networks 54, 2787–2805.

Ballagas, R., Borchers, J., Rohs, M., Sheridan, J.G., 2006. The smart phone: a ubiquitous input device. IEEE Pervasive Comput. 5, 70–77.

Bao, L., Intille, S.S., 2004. Activity Recognition from User-Annotated Acceleration Data, in: Ferscha, A., Mattern, F. (Eds.), Pervasive Computing, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 1–17.

Beresford, A., Stajano, F., 2003. Location privacy in pervasive computing. IEEE Pervasive Comput. 2, 46–55.

Bouten, C.V.C., Koekkoek, K.T.M., Verduin, M., Kodde, R., Janssen, J.D., 1997. A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. IEEE Trans. Biomed. Eng. 44, 136–147.

Chaudhary, N., 2013. An open-source framework for context-aware monitoring of mobile application feature usage. Aalto University Master's Thesis.

Figo, D., Diniz, P.C., Ferreira, D.R., Cardoso, J.M.P., 2010. Preprocessing techniques for context recognition from accelerometer data. Personal Ubiquitous Computing 14, 645–662.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm. J. R. Stat. Soc. Ser. C Appl. Stat. 28, 100–108.

He, J., Li, H., Tan, J., 2007. Real-time Daily Activity Classification with Wireless Sensor Networks using Hidden Markov Model, in: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007. Presented at the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007, pp. 3192–3195.

Huynh, T., Schiele, B., 2005. Analyzing Features for Activity Recognition, in: Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-Aware Services: Usages and Technologies, sOc-EUSAI '05. ACM, New York, NY, USA, pp. 159–163.

Iftode, L., Borcea, C., Ravi, N., Kang, P., Zhou, P., 2004. Smart Phone: an embedded system for universal interactions, in: 10th IEEE International Workshop on Future Trends of Distributed Computing Systems, 2004. FTDCS 2004. Proceedings of the 10th IEEE International Workshop on Future Trends of Distributed Computing Systems, 2004. FTDCS 2004. pp. 88–94.

Jain, A.K., 2010. Data clustering: 50 years beyond K-meansq. Pattern Recognition Letters 31, 651–666.

Kantola, J., Perttunen, M., Leppänen, T., Collin, J., Riekki, J., 2010. Context awareness for gps-enabled phones, in: Proceedings of ION Technical Meeting. pp. 117–124.

Kwapisz, J.R., Weiss, G.M., Moore, S.A., 2011. Activity Recognition Using Cell Phone Accelerometers. SIGKDD Explor Newsl 12, 74–82.

Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T., 2010. A survey of mobile phone sensing. IEEE Commun. Mag. 48, 140–150.

Lee, Y.-S., Cho, S.-B., 2011. Activity Recognition Using Hierarchical Hidden Markov Models on a Smartphone with 3D Accelerometer, in: Corchado, E., Kurzyński, M., Woźniak, M. (Eds.), Hybrid Artificial Intelligent Systems, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 460–467.

Long, X., Yin, B., Aarts, R.M., 2009. Single-accelerometer-based daily physical activity classification, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009. Presented at the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009, pp. 6107–6110.

Mannonen, P., Karhu, K., Heiskala, M., 2013. An Approach for Understanding Personal Mobile Ecosystem in Everyday Context. Eff. Agile Trust. Eser. Co-Creat. pp. 135-146.

Phithakkitnukoon, S., Horanont, T., Lorenzo, G.D., Shibasaki, R., Ratti, C., 2010. Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data, in: Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (Eds.), Human Behavior Understanding, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 14–25.

Randell, C., Muller, H., 2000. Context awareness by analysing accelerometer data, in: The Fourth International Symposium on Wearable Computers. Presented at the The Fourth International Symposium on Wearable Computers, pp. 175–176.

Ravi, N., Dandekar, N., Mysore, P., Littman, M.L., 2005. Activity recognition from accelerometer data, in: AAAI. pp. 1541–1546.

Reddy, S., Burke, J., Estrin, D., Hansen, M., Srivastava, M., 2008. Determining transportation mode on mobile phones, in: 12th IEEE International Symposium on Wearable Computers, 2008. ISWC 2008. Presented at the 12th IEEE International Symposium on Wearable Computers, 2008. ISWC 2008, pp. 25–28.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M., 2010. Using mobile phones to determine transportation modes. ACM Trans Sen Netw 6, 13:1–13:27.

Swan, M., 2012. Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. J. Sens. Actuator Netw. 1, 217–253.

Swan, M., 2013. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. Big Data 1, 85–99.

Ward, J.A., Lukowicz, P., Troster, G., Starner, T.E., 2006. Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers. IEEE Trans. Pattern Anal. Mach. Intell. 28, 1553–1567. doi:10.1109/TPAMI.2006.197

Yang, J.-Y., Chen, Y.-P., Lee, G.-Y., Liou, S.-N., Wang, J.-S., 2007. Activity Recognition Using One Triaxial Accelerometer: A Neuro-fuzzy Classifier with Feature Reduction, in: Ma, L., Rauterberg, M., Nakatsu, R. (Eds.), Entertainment Computing – ICEC 2007, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 395–400.

Yi, J.S., Choi, Y.S., Jacko, J.A., Sears, A., 2005. Context Awareness via a Single Device-attached Accelerometer During Mobile Computing, in: Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices &Amp; Services, MobileHCI '05. ACM, New York, NY, USA, pp. 303–306.

Zhang, S., McCullagh, P., Nugent, C., Zheng, H., 2010. Activity Monitoring Using a Smart Phone's Accelerometer with Hierarchical Classification, in: 2010 Sixth International Conference on Intelligent Environments (IE). Presented at the 2010 Sixth International Conference on Intelligent Environments (IE), pp. 158–163.